Non-Gaussian, Non-stationary and Nonlinear Signal Processing Methods - with Applications to Speech Processing and Channel Estimation

Li, Chunjian

Publication date:
2007

Document Version
Early version, also known as pre-print

Non-Gaussian, Non-stationary, and Nonlinear Signal Processing Methods

with Applications to Speech Processing and Channel Estimation

Ph.D. Thesis

CHUNJIAN LI

February, 2006

Chunjian Li
Non-Gaussian, Non-stationary, and Nonlinear Signal Processing methods - with Applications to Speech Processing and Channel Estimation

ISBN 87-90834-90-9
ISSN 0908-1224
R06-1001

Department of Communication Technology
Aalborg University
Fredrik Bajers Vej 7
DK-9220 Aalborg Øst, Denmark
Phone: +45 9635 8650

This thesis was typeset using LaTeX.

Printed by Uniprint, Aalborg, Denmark.

Abstract

The Gaussian statistic model, despite its mathematical elegance, is found to be too factitious for many real world signals, as manifested by its unsatisfactory performance when applied to non-Gaussian signals. Traditional non-Gaussian signal processing techniques, on the other hand, are usually associated with high complexities and low data efficiencies. This thesis addresses the problem of optimum estimation of non-Gaussian signals in computation-efficient and data-efficient ways. The approaches that we have taken exploit the high temporal-resolution non-stationarity or the underlying dynamics of the signals. The sub-topics being treated include: joint MMSE estimation of the signal DTFT magnitude and phase, high temporal-resolution Kalman filtering, blind de-convolution and blind system identification, and optimum non-linear estimation. Applications of the proposed algorithms to speech enhancement, non-Gaussian spectral analysis, noise-robust spectrum estimation, and blind channel equalization are demonstrated.

The thesis consists of two parts, the Introduction and the Papers. The Introduction gives background information of the problems at hand, states the motivation of approaches taken, summarizes the state-of-the-art in literature, and describes our contributions briefly. The Papers presents our contributions in the form of published papers.

The first part of the Papers (paper A and B) deals with the importance of phase in non-Gaussian signal estimation. Joint MMSE estimators of both magnitude spectra and phase spectra are developed. Application to the enhancement of noisy speech signals results in clearer sounds and higher SNR than frequency domain MMSE estimators. Here the non-Gaussianity of the speech signal is modeled by the linearity in the phase spectrum, and is enhanced by the joint estimator. This is in contrast to the spectral domain MMSE estimator (e.g., the Wiener filter), which is zero-phase.

The second part of the Papers (paper C and D) attacks the non-Gaussian estimation problem with a purely temporal domain approach. It is recognized that a temporal-domain high-resolution non-stationary LMMSE estimator is able to extract structures in both magnitude and phase spectra at a lower complexity. For speech signals, the non-Gaussianity is represented by an excitation sequence with a rapidly varying vari-

ance filtered by an all-pole filter. A Kalman filter with a time-varying system noise is ideally suitable to this model. This so called high temporal-resolution Kalman filtering technique fully exploits the non-stationary processing capability of the Kalman filter, yet takes advantage of the fact that the all-pole filter changes slowly over time. This is in contrast to the conventional frame-based Kalman filtering, which presumes signals to be stationary within a processing frame, and to the adaptive Kalman filtering which adapts all system parameters in every time instant.

The third part of the Papers (paper E, F and G) sees the non-Gaussian estimation problem from yet another angle. Her the non-Gaussian excitation is treated as a discrete-state finite-alphabet symbol sequence. The new model combines the HMM and the AR model to represent a wide range of signals, thus we call it the Hidden Markov-Autoregressive model (HMARM). The HMARM can efficiently extract the second order and higher order temporal structure with the two dynamic models respectively. Efficient ML system identification algorithms are derived based on the EM methodology to jointly estimate the HMM parameters and the AR parameters. In paper F, the HMARM is extended to having a measurement noise at the output of the AR model. This extension increases the estimation complexity significantly since the system output is now hidden and the measurement noise variance need to be estimated jointly with other parameters. A nonlinear MMSE estimator is incorporated into the EM algorithm to provide the sufficient statistics for the learning. The HMARM and its extended version are applied to speech analysis, noise robust spectrum estimation, and blind channel equalization for PAM and PPM signals.

The proposed algorithms in this thesis only involve computations of the second order statistics explicitly. The higher order structure is though represented by the appropriately chosen models. Thus the computational complexity is low and data efficiency is high compared to Higher Order Statistics based methods, which require no signal models.

List of Papers

The thesis is based on the following papers:

[A] Chunjian Li and Søren Vang Andersen, "Inter-frequency Dependency in MMSE Speech Enhancement". In *Proceedings of the 6th Nordic Signal Processing Symposium, NORSIG-2004*, pp. 200-203. June 9-11, 2004, Espoo, Finland.

[B] Chunjian Li and Søren Vang Andersen, "A Block Based Linear MMSE Noise Reduction with a High Temporal Resolution Modeling of the Speech Excitation". In *EURASIP Journal on Applied Signal Processing, Special Issue on DSP in Hearing Aids and Cochlear Implants*, vol. 2005:18, pp. 2965-2978. October, 2005.

[C] Chunjian Li and Søren Vang Andersen, "Integrating Kalman filtering and multipulse coding for speech enhancement with a non-stationary model of the speech signal". In *Proceedings of the Thirty-eighth Annual Asilomar Conference on Signals, Systems, and Computers, ASILOMAR-2004*. November 7-11, 2004. Pacific Grove, California, USA.

[D] Chunjian Li and Søren Vang Andersen, "A new Iterative Speech Enhancement Scheme Based on Kalman Filtering". In *Proceedings of the 13th European Signal Processing Conference, EUSIPCO-2005*. September 9-11, 2005, Antalya, Turkey.

[E] Chunjian Li and Søren Vang Andersen, "Blind Identification of Non-Gaussian Auto-regressive Models for Efficient Analysis of Speech Signals". In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. May 14-19, 2006, Toulouse, France.

[F] Chunjian Li and Søren Vang Andersen, “Efficient Blind System Identification of Non-Gaussian Auto-Regressive Models with Dynamic Modeling”. Accepted for publication in *IEEE Transactions on Signal Processing*.

[G] Chunjian Li and Søren Vang Andersen, “Efficient Implementation of the HMARM Identification and Its Application in Spectral Analysis”. Submitted to *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2007*.

The research that is documented in this thesis has lead to the provisional filing of the following patents:

[1] Chunjian Li and Søren Vang Andersen. A method for noise reduction using a non-Toeplitz temporal signal covariance matrix, 2005.

[2] Chunjian Li and Søren Vang Andersen. Efficient initialization of iterative parameter estimation, 2005.

[3] Chunjian Li and Søren Vang Andersen. High temporal resolution estimation of LPC excitation variance of signals, 2005.

[4] Chunjian Li and Søren Vang Andersen. A non-Gaussian signal analysis technique, 2006.

Preface

This thesis is submitted to the International Doctoral School of Technology and Science at Aalborg University as a partial fulfillment of the requirements for the degree of Doctor of Philosophy. The work was carried out during the period March 1st, 2003 - February 28th, 2006 at the Department of Communication Technology at Aalborg University, and was funded by The Danish National Centre for IT Research, Grant No. 329 and Microsound A/S.

I would like to thank my primary supervisor Søren Vang Andersen for his professional guidance and constant support. His encouragement has always strengthened me when I explored new ideas, and his broad knowledge has been an important source of my learning and building my own competence. I would also like to thank my co-supervisors Søren Holdt Jensen, Per Rubak, and Uwe Hartmann for many fruitful discussions and valuable advices. I also appreciate the effort of Søren Louis Petersen at Microsound A/S and Kjeld Hermansen in making the project a reality. I also thank them and other Microsound employees who have involved in the project for bringing in their industrial viewpoints and technical contributions, from which I have gotten inspirations and insights.

Finally, I would like to acknowledge my colleagues and fellow Ph.D. students at Aalborg University. Special thanks go to Karsten Vandborg Sørensen, who I have shared room and project with for the past three years, for giving me assistance that makes my stay in Denmark easier; Mads Græsbøll Christensen, Xuefeng Yin, Morten Holm Larsen, Steffen Præstholm, Ingmar Land, Troels Pedersen, Joachim Dahl, Bin Hu, and Christoffer Asgaard Rødbro for many interesting discussions. Last, but not least, I thank my friends and family for their support during my time at Aalborg University.

Chunjian Li
Aalborg, February 2006

Acronyms

AR	auto-regressive
ARMA	auto-regressive moving average
ARX	auto-regressive with exogenous input
BLUE	best linear unbiased estimator
E-HMARM	extended hidden Markov auto-regressive model
DFT	discrete Fourier transform
EKF	extended Kalman filter
EM	expectation-maximization algorithm
GEM	generalized EM algorithm
GMM	Gaussian mixture model
GSF	Gaussian sum filter
HMARM	hidden Markov auto-regressive model
HMM	hidden Markov model
HOS	higher order statistics
i.i.d.	independent and identically distributed
ISI	inter-symbol interference
KF	Kalman filter
LDA	Levinson-Durbin algorithm
LMMSE	linear minimum mean square error
LPC	linear predictive coding
LS	least squares
LTI	linear time-invariant
MAP	maximum a posterior
ML	maximum likelihood
MP	matching pursuit
MPLPC	multi-pulse linear predictive coding
MMSE	minimum mean squared error
MSE	mean square error
PAM	pulse amplitude modulation

PDF	probability density function
PPM	pulse position modulation
SKF	switching Kalman filter
SNR	signal to noise ratio
SVD	Singular Value Decomposition
TLS	total least squares
XLS	extended least squares
WF	Wiener filter
WSS	wide sense stationary

Contents

Part I

Introduction

Introduction

1 Non-Gaussian time series and Bayesian estimation

A time series is a sequence of observations that are orderly in time (or space). Most of the natural and man-made signals are time series, e.g. speech, images, and communication signals. Many important time series exhibit certain temporal structures, or temporal dependencies. Temporal dependency in a time series is often modeled by linear models, such as auto-regressive (AR), moving average (MA), and autoregressive-moving average (ARMA) models, although nonlinear temporal dependency is sometimes of interest and can be modeled by nonlinear models such as the Volterra series [1] [2] and neural network based models [3]. A linear model can be seen as a linear time invariant (LTI) filter excited by a stationary Gaussian process, whereas a nonlinear model can be seen as a nonlinear filter excited by either a Gaussian or a non-Gaussian process. In this work, we focus on LTI filters, especially the AR filters, excited by non-stationary or non-Gaussian processes. The motivation is that linear filters are easier to analyze, and, as will be shown later on, the LTI filter model with a non-stationary/non-Gaussian input is able to represent a wide range of nonlinear signals.

In the category of linear models, the AR model is the most frequently used in applications. There are several reasons for its popularity: 1) the AR model can well represent spectra with narrow peaks, and narrow band spectra are very common in practice [4]; 2) for a Gaussian process, the maximum entropy spectrum [5] is the spectrum given by AR modeling [6]; 3) under the Gaussian assumption, the AR parameter estimation problem is linear while the MA and ARMA estimation problems are nonlinear. Moreover, the AR model with a sufficiently high order can be used to approximate any ARMA models arbitrarily well [7, p.52] [8, p.411].

Under the standard definition of the AR model, an AR process is created by filtering an independent, identically distributed (i.i.d.) sequence by an all-pole filter [9] [10]. The most used distribution in the AR modeling is the Gaussian pdf. This model is, however, too restrictive to suit many important signals. As we will show later, voiced speech signals and some communication signals are better modeled having non-Gaussian or non-

i.i.d. processes as inputs to the all-pole filters. In this thesis, we use a generalized AR model definition in which the input process to the all-pole filter can be non-Gaussian, non-stationary, and temporally dependent.

Definition 1 *The process $\{\mathbf{X}_t\}$ is said to be a generalized AR(p) process if for every t it satisfies the difference equation*

$$\mathbf{X}_t - a_1\mathbf{X}_{t-1} - \cdots - a_p\mathbf{X}_{t-p} = \mathbf{Z}_t, \tag{1}$$

where $\mathbf{Z}_t$ is a random process that can take on any probability density function (pdf), can be non-stationary within the analysis frame, and can be temporally dependent.

Remark 1: The generalized AR model belongs to the big category of equation-error-type models, which is defined in [11, p.71, p.74]. All the AR models mentioned in the sequel are under this generalized definition.

Remark 2: This definition means that the input process $\mathbf{Z}_t$ can be any time series. This is especially useful for de-convolution problems.

When the excitation process $\mathbf{Z}_t$ in an AR model is stationary, white, and Gaussian, the model is known as the Gaussian AR model. The Gaussian AR model has been widely used in many signal processing fields including linear prediction [12] [13], spectral analysis [6] [14], and linear dynamical modeling [15, p.420] [16]. The identification of the Gaussian AR model has also been extensively studied. Thanks to the stationary-white-Gaussian assumption, the Gaussian AR parameters can be identified analytically using, e.g. the Least Squares (LS) method [11] [15] [4].

When the excitation process $\mathbf{Z}_t$ is i.i.d. non-Gaussian, the model is known as the non-Gaussian AR model. Non-Gaussian AR models have recently attracted an increased attention in the signal processing society. Many signals are found to be far from Gaussian [17] [18] [19]. In other words, for many signals, non-Gaussian stochastic models often outperform Gaussian models significantly and can be used to solve problems that are unsolvable with the Gaussian models (e.g. Blind Source Separation using Independent Component Analysis [20]). Major benefits of non-Gaussian estimation includes smaller estimation variance and bias [21] [22], robustness to outliers [23], and efficient representation of signals [23] [24] [25]. Research works on non-Gaussian AR modeling have appeared in image processing [26] [27] [28], speech processing [29] [23], medical signal processing [30], radar signals [31], navigation [32], econometrics [33], and communications signal processing [34].

When the excitation process $\mathbf{Z}_t$ is a non-stationary Gaussian process with possibly temporal dependency, i.e., a non-i.i.d.[1] Gaussian process, it is often treated as an i.i.d. non-Gaussian process too. Note that here, we are talking about a Gaussian process that

[1]Here, a non-i.i.d. process is referred to as a non-independent and/or non-identically distributed random process.

changes its mean and/or variance at every time instance, such that the usual short-time processing techniques (based on the quasi-stationary assumption) are not applicable.

Similar generalizations of the linear time-invariant (LTI) system to accommodating non-Gaussian input process date back to the 60's. Bartlett [35] in 1955 and Brillinger et al. [36] in 1967 analyzed the polyspectra for the i.i.d. non-Gaussian and non-i.i.d. processes excited linear systems (see [37]). In [11], ARMA models are generalized such that the modeling errors are themselves AR or MA processes, therefore correlated errors are introduced. In [38, Theorem 2], it is shown that a linear system with a non-i.i.d., non-Gaussian input process can be identified using higher order statistics. The non-i.i.d. Gaussian excited AR process, though, has received less research attention than the i.i.d. non-Gaussian excited AR process. In this work, we promote the use of the non-i.i.d. Gaussian excited AR process, and we give the following motivations for it: 1) its optimum filtering problem can be solved analytically, with appropriate adaptations to the classical optimum linear filters; 2) there is often rich temporal structures in the input process which can be exploited to facilitate the identification of the underlying dynamics of the non-stationary Gaussian process, while the i.i.d. non-Gaussian model ignores this temporal structure.

It is well known that a nonlinear transformation of an i.i.d. Gaussian process in general results in an i.i.d. non-Gaussian process. We contend here that a non-stationary, though linear, transformation of a Gaussian process can also make an i.i.d. non-Gaussian distribution if viewed as a static system. By non-stationary linear transform, we mean the transform that changes its functional form or coefficients along time. As an example,

$$Y = a_t X + b_t \tag{2}$$

is such a transform, where X is a stationary Gaussian process, a_t and b_t are the transform coefficients that change over time. The resulting process Y can be seen as either a non-Gaussian process if assumed stationary, or a non-stationary process if assumed Gaussian. In other words, the same set of data can be explained by either a statistical structure in a static view, or a temporal structure in a dynamical view. Fig. 1 shows the relations between the two transforms. The double-arrow in the center shows the duality, i.e., a process can be modeled as an i.i.d. non-Gaussian process by ignoring the temporal structure in it, or modeled by a non-i.i.d. Gaussian process if the temporal structure can be identified.

We prefer to use the dynamical view anywhere possible, since it allows analytical solution to the optimum estimation problem now that the Gaussian assumption is maintained. Such observations are analogous to the time-variant linear system theory, which linearizes a nonlinear system along its trajectory and results in a time-variant linear system. The Extended Kalman filter (EKF) [39] is a good example of such a dynamical linearization. But unlike the EKF, the non-i.i.d. Gaussian AR model confines its

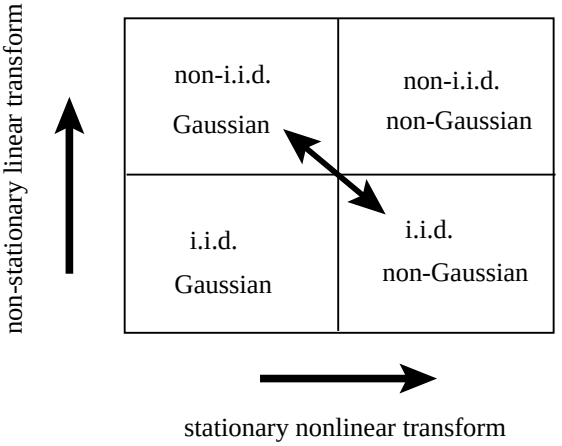


Figure 1: Non-Gaussianity, non-stationarity, and nonlinearity.

nonlinearity in the input process instead of the filter. This brings several benefits:

1. The filter is linear and is easier to identify;
2. The nonlinearity of the input process is in the form of a non-Gaussian pdf, which has no problem of representing discontinuity such as switching effects. Whereas the EKF requires the existence of derivatives of the nonlinear function.
3. This is useful in many de-convolution problems, where the input to the filter has non-Gaussian structures.

The applicability of the dynamical view, however, requires knowledge of the dynamics of the input process. For example, in [18, p.145], a switching model in which one of its constituent Gaussian sub-processes is selected at each instant is shown to have a non-Gaussian pdf, since its switching is random. A switching process can not be treated as a non-stationary Gaussian unless the switching is deterministic. In other words, if the switching mechanism is decoded, the switching process can be modeled by a non-stationary Gaussian process without losing any information.

We are interested in two types of non-stationary Gaussian input processes: the Gaussian process with a time-varying variance, and the Gaussian process with a time-varying mean. In contrast to the conventional AR model whose input process must be white, there can be temporal dependency in the input process of the generalized AR model. In fact, temporal dependency in the input process is welcomed in our models since it facilitates the estimation of the temporal structure. An example of the non-stationary-in-variance Gaussian process with temporal dependency is a Gaussian process with a smoothly varying variance. An example of the non-stationary-in-mean process with

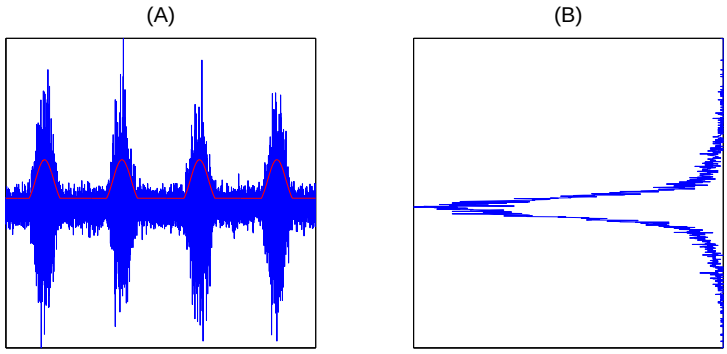


Figure 2: (A) A non-stationary Gaussian process with a smoothly varying variance. The red curve is the scaling factor as a function of time. (B) The resulting histogram is non-Gaussian.

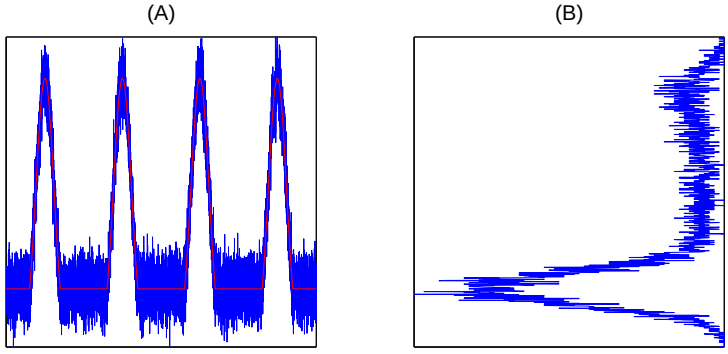


Figure 3: (A) A non-stationary Gaussian process with a smoothly varying mean. The red curve is the mean as a function of time. (B) The resulting histogram is non-Gaussian.

temporal dependency is a Gaussian process with a smoothly varying mean. An example of the switching process with deterministic switching is a GMM or HMM process with decoded states. Fig. 2 and Fig. 3 shows examples of non-Gaussian processes created by varying the variance or mean of a Gaussian process, and Fig. 4 shows a switching process with two Gaussian components. They all can alternatively be seen as non-Gaussian processes if viewed statically (by the histograms).

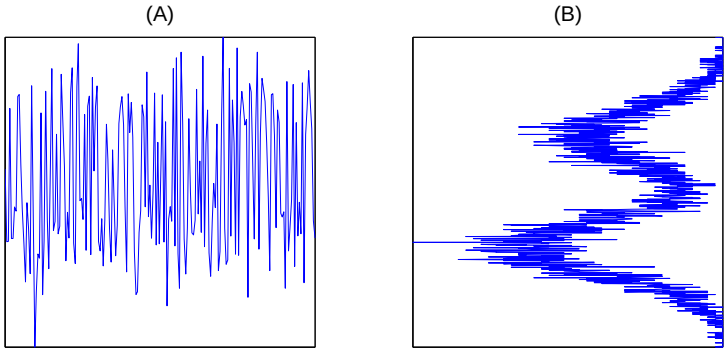


Figure 4: (A) A switching process with deterministic switching states. (B) The resulting histogram is non-Gaussian.

Bayesian estimation of non-Gaussian signals

Despite the promising results given by non-Gaussian signal processing techniques, theories and methods in this field are still underdeveloped. Fundamental problems such as optimum filtering of non-Gaussian signals and parameter estimation of non-Gaussian models are still difficult. The major difficulty is that optimum non-Gaussian estimation problems are nonlinear. So either a nonlinear equation system needs to be solved (in estimating parameters), or numerical integration of an arbitrary pdf need to be evaluated (in filtering). These problems become even more difficult when the signal is a non-Gaussian AR process instead of a non-Gaussian i.i.d. process, because the pdf of the non-Gaussian AR process evolves along time axis, unlike the stable pdf in the i.i.d. case.

Recognizing the difficulty of the general non-Gaussian signal processing problem, we, in this thesis, avoid solving the problem in a general sense. Instead, we attack the problem by taking on a particular type of signals that have powerful structures which can be exploited for efficient filtering and system identification. This class of signals are the generalized AR signals with prominent temporal structures in their input processes. The signals that we treated in this thesis include voiced speech signals, Pulse Amplitude Modulation (PAM) signals and Pulse Position Modulation (PPM) signals with Inter-Symbol Interference (ISI). A wide range of other signals are suitable for this model too, although not treated in this work, such as images, music, and radar signals.

Here we define the signal estimation process as the act of recovering a signal waveform from its distorted or noisy observations. Any time series estimation problem can be decomposed into three basic tasks: model design, estimation of model parameters, and estimation of the time series given the estimated model. In statistics and neural

networks literature, the last two tasks are also known as learning (of the model) and inference (of the data). These terms will be used interchangeably in the sequel.

In this work, we consider Bayesian estimation methods, in particular the Minimum Mean Squared Error (MMSE) estimator, for the signal estimation problem. Bayesian estimation provides a convenient framework for exploiting prior knowledge of the signal statistics in the estimation. The prior knowledge is represented by the prior probability distribution. For a Gaussian AR process, the prior is a Gaussian pdf, while for a non-Gaussian AR process the prior takes the form of a non-Gaussian pdf. It is well known that the Bayesian methods result in linear estimators only if the signals are Gaussian. For arbitrary priors, the Bayesian estimators are generally nonlinear.

Established methods for solving non-Gaussian MMSE estimation problems can be grouped as follows:

1. integrating non-Gaussian parametric pdfs, which results in highly nonlinear equations [40] [41];
2. approximating priors using Gaussian Mixture Models (GMM), which results in the Gaussian Sum Estimator [42] [43];
3. using sampling techniques to approximate the pdf, which results in the Monte Carlo filters [44] [45] [46] [47] [48].

The problem with the first group of methods is that, the closed form nonlinear solutions do not generally exist. Even the proposed ones are obtained under very restrictive assumptions. For the Gaussian Sum Estimator, a major drawback is that the number of constituent states grows exponentially with the time index, and so does the complexity. The Monte Carlo filters are also associated with high complexities since large numbers of samples need to be generated and their likelihood to be tested.

In the works included in this thesis, we adapt a general strategy different from the above. Specifically, we extend the classic linear Gaussian models to accommodate non-Gaussian signals by exploiting special temporal structures in the signals. In this way, the complexity is maintained at a comparable level with the linear Gaussian methods, while the non-Gaussian features of the signals are faithfully represented. In the following sections, the signal structures of interest are first introduced, then classic methods in Bayesian signal estimation and parameter estimation will be briefly reviewed, and our views on how these problems should be approached in the non-Gaussian case will be briefly introduced.

2 Temporal structures of non-Gaussian AR signals

A time series carries information in its temporal structure, eg. audio signals, images, and certain modulated signals used in communications, just to name a few. This is

in contrast with signals that carry information in the frequency of occurrence, eg. the failure rate of a component, the bit-error rate of a communication system, results of independent experiments, the histogram of a random process, and etc. Thus in time series modeling, exploiting temporal structure is one of the key factors. Here the temporal structure is defined as any pattern exhibited by the signal in the time domain that can be described by a mathematical model with a small number of coefficients. The conventional Gaussian AR model however,

1. only models the signal correlation, which is a second order dependency;
2. contributes all signal correlation to the all-pole filter, even though for some AR signals the input processes are not white.

Many signals have prominent temporal structures in the input process when modeled by the AR model. In this work, we study two important groups of signals: speech signals and communications signals.

Specifically, the speech signals that are of interest here are the voiced speech signals, and the communication signals that are of interest are the PAM and PPM signals with ISI. When modeled by the AR model, the residual of the voiced speech signal exhibits an impulse train structure, as shown in Fig. 5. This structure has long been recognized to be important to the speech quality in speech coding literature [49] [50]. In the filtering problem, this structure is usually ignored due to the use of linear Gaussian models. To exploit this structure, from a Bayesian optimum filtering point of view, the input process can be modeled by a super-Gaussian pdf (e.g., Laplace distribution) [51] [52], due to the large amplitude of the spikes. Solving for the MMSE estimate requires integrating the non-Gaussian pdf, which is generally intractable for high-dimension problems. In the first part of the Papers, We propose to model the input process as a non-stationary Gaussian process with a constant mean and a time-dependent variance. The variance goes up at the vicinity of an impulse and remains low between the impulses. Thus, the time-dependent variance can represent the temporal localization of the power in the input process. As will be shown below, this high temporal resolution modeling brings in many advantages for both the block-based spectral domain MMSE estimator and the temporal domain sequential MMSE estimator.

In the second part of the Papers, we propose to model the input process as a sequence of discrete-valued symbols from a finite alphabet added with white Gaussian noise. A Hidden Markov Model (HMM) is ideal for modeling such a process, with the assumption that the temporal dependency is Markovian. The HMM can be seen as a Kalman filter model with a simple nonlinearity [53]. It can also be seen as modeling a Gaussian process with a mean controlled by a switching mechanism that is nonlinear. More about the HMM and nonlinear filtering will be introduced in Section 3.3. When the HMM is cascaded with the AR model, they respectively extract the nonlinear temporal dependency and the linear dependency from the signal. This model can represent

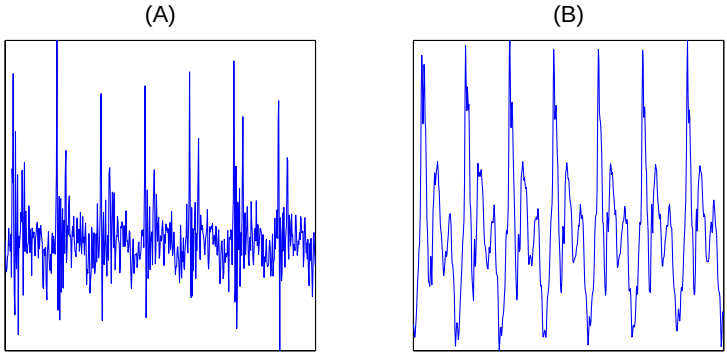


Figure 5: (A) LPC residual of the vowel /ae/. (B) The waveform of the speech.

a broader range of signals that have equivalent discrete input processes with temporal dependency. Besides the analysis of voiced speech signals, we have investigated the channel equalization problem of PAM and PPM signals. Specifically, the ISI channel is modeled as an AR filter, with or without additive measurement noise, and the transmitted symbols are modeled by the HMM. If the transmitted sequence of symbols possess a certain dependency, the HMM can capture it and exploit it in the filtering. The dependency between symbols is due to the special way the symbols are arranged, such as the PPM signals, or is introduced into the sequence on purpose, such as the trellis modulated signals [54]. If the transmitted symbols are indeed i.i.d., such as ordinary PAM signals, the HMM reduces to a Gaussian Mixture Model (GMM). Fig. 6 shows an example of PPM signals.

3 Signal estimation

This section reviews the estimation of the signal waveform of an AR(p) process, assuming that the signal model and its parameters are known. For an AR(p) process we have the following signal model

$$x(t) = \sum_{k=1}^{p} a_k x(t-k) + u(t), \tag{3}$$

$$y(t) = x(t) + v(t), \tag{4}$$

where $y(t)$ is the observation, $x(t)$ is the clean signal, $v(t)$ is the observation noise, $u(t)$ is the excitation process to the AR(p) filter, and a_k are the AR coefficients. The signal

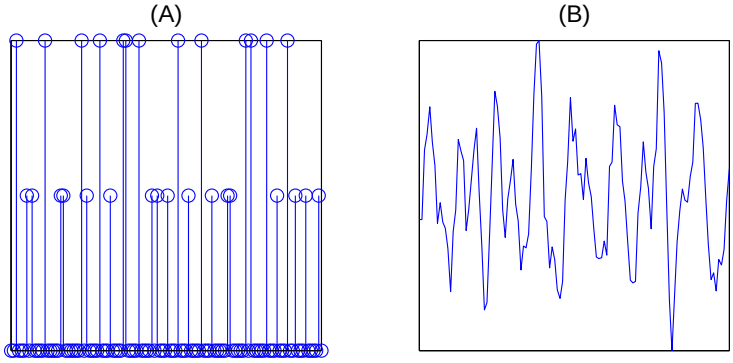


Figure 6: (A) The transmitted symbol sequence of a combined PPM-PAM modulation. (B) The received waveform, assuming the channel is AR(10).

model (3) and (4) are also known as the linear dynamic model.

To simplify the presentation, we assume that the noise $v(t)$ is an i.i.d. Gaussian process. In the case that the additive noise is correlated in time or non-Gaussian, the noise should be treated as another signal, and optimum joint estimation of the two signals can be done by generalizing the estimator to its vector form. This is more of a topic of source separation, which is not addressed in this thesis.

3.1 Wiener filtering

The causal Wiener filter (WF) is a Linear Minimum Mean Squared Error (LMMSE) estimator of the signal $x(t)$ given the observation $y(k)$ for $-\infty < k \leqslant t$. The causal Wiener filter is rarely used in practice due to the difficulty of a required spectral factorization procedure [55, p.265]. Commonly used in practice is the non-causal Wiener filter (or Wiener smoother). We will now review both filters.

Causal Wiener filters

The LMMSE estimator solves a special case of the MMSE estimation problem, in which the priors of the clean signal and the observation noise are assumed to be Gaussian. We use the Gaussian AR signal model (3) and (4) again. To be convenient, we re-write the signal model in matrix form.

$$\mathbf{y} = \mathbf{x} + \mathbf{v}, \tag{5}$$

where the boldface letters represent N dimensional vectors that contain the data from time 1 to N. The LMMSE estimate of the signal $\mathbf{x}$ can be shown to be the conditional

expectation of the signal given the observation $\mathbf{y}$ [15]:

$$\begin{aligned}\hat{\mathbf{x}} &= E[\mathbf{x}|\mathbf{y}] \\ &= \mathbf{C_{xy}}\mathbf{C_{yy}^{-1}}\mathbf{y},\end{aligned} \tag{6}$$

where $\mathbf{C_{yy}}$ is the $N \times N$ covariance matrix of $\mathbf{y}$, and $\mathbf{C_{xy}}$ is the $N \times N$ cross-covariance matrix of $\mathbf{x}$ and $\mathbf{y}$. In practical problems, the covariance matrix of the clean signal is unknown and difficult to estimate. In the Wiener theory, the signal length N is assumed to be infinitely long, spanning from time $-\infty$ to present time. Based on this assumption and the stationarity assumption, Wiener and Hopf proposed a spectral factorization method to find the spectral response of the causal Wiener filter using power spectral density (psd) of the signal, which is much easier to estimate than the covariance matrix [55, p.231] [10, p.417]. Notice that in this method, the signal is assumed to be wide sense stationary (WSS) in order to use the power spectral density, and the signal length is assumed to be semi-infinite.

Non-causal Wiener filters

The non-causal Wiener filter solves the problem by assuming the signal length and the filter taps length to be infinite, in addition to the WSS assumption. Now, to minimize the MSE of the estimate by applying the orthogonality principle, one obtains the following equation:

$$R_{yx}(t) = \sum_{k=-\infty}^{+\infty} h(k) R_{yy}(t-k) \qquad \text{for all t,} \tag{7}$$

where $h(k)$ is the kth coefficient of the Wiener filter, $R_{yx}(t)$ is the cross-correlation function of the $y(t)$ and $x(t)$, $R_{yy}(t)$ is the auto-correlation function of $y(t)$. Because of the infinite summation, taking the Fourier transform of both sides of (7) results in

$$S_{yx}(f) = H(f) S_{yy}(f), \tag{8}$$

or

$$H(f) = \frac{S_{yx}(f)}{S_{yy}(f)}, \tag{9}$$

where $S_{yx}(f)$ and $S_{yy}(f)$ are the psds, and $H(f)$ is the frequency response of the Wiener filter.

Extension to the Wiener filter

In both the causal and non-causal Wiener filter, it is assumed that the signal is wide sense stationary and the signal length is infinite or semi-infinite. These assumptions are obvi-

ously inappropriate in practical problems. First, the observation data are often of short length. Short time processing is a common technique in many signal processing applications, such as speech processing. When the length of the data frame is comparable to the correlation span of the signal, the stationarity assumption does not hold. Second, the local stationarity assumption rules out the possibility of modeling the dynamics of the signal within the processing frame. For a time series that has rich temporal structures, the stationarity assumption is a major drawback. As consequences, the Wiener filter 1) provides only trivial estimate of the phase spectrum; 2) does not exploit potential inter-frequency correlation; 3) does not suppress noise power according to temporal distribution of the signal power.

As an example, we consider the voiced speech signal. A frame of voiced speech can be modeled by filtering a noisy impulse train by an AR filter. This is known as the speech production model, or the source-filter model and is widely used in speech coding and speech synthesis [49]. It is obviously a non-Gaussian AR model, since the input to the AR filter is super-Gaussian due to the large values of the impulses. Because of the mechanism of glottal folds movement, the excitation to the AR filter has an impulse train structure. Instead of modeling this temporal structure with a static super-Gaussian model, it is beneficial to model it as a non-stationary Gaussian process with rapidly varying variance. That is, between two impulses, the process has a low variance, and at the vicinities of the impulses, the process has large variances. The large variance represent the concentration of power at certain time points.

We show in paper A and B, that with a high temporal resolution modeling of the input process, a block based LMMSE estimator can be obtained, which jointly estimates the phase and magnitude spectra of the signal, exploits inter-frequency correlation to help estimation of those spectral components with low local SNRs, and attenuates noise power at the valleys between the excitation impulses.

Frequency domain methods

In the speech processing literature, estimation methods based on frequency domain manipulations are dominant, e.g. the power spectral subtraction method [56], the MMSE short-time spectral amplitude estimator [40], the MAP spectral amplitude estimator and MMSE spectral power estimator [57], and the MMSE estimator of magnitude-squared DFT coefficients [41]. These estimators only estimate the spectral magnitude and have zero phase, and they all assume stationarity of the signal and independence between spectral components. Thus they share the same property of the non-causal Wiener filter as discussed above.

3.2 Kalman filtering

The Kalman filter is a very important extension to the Wiener filter within the LMMSE framework. The Kalman filter generalizes the LMMSE estimator to allow the parameters to evolve in time. This is possible because of the use of the state-space model and sequential estimation. Thanks to its capability of handling non-stationary signals, the Kalman filter is ideal for our high temporal resolution modeling of the input process to the AR filter. Also, because the Kalman filter is a time domain method, it has no such problem as ignoring phase spectra as in the Wiener filter (Wiener filter is sometimes referred to as a time domain method, whereas it is indeed solved in the spectral domain).

The Gaussian AR signal model (3) and (4) can be written in the standard state-space form:

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{A}\mathbf{x}(n-1) + \mathbf{b}u(n) \\ y(n) &= \mathbf{h}\mathbf{x}(n) + v(n), \end{aligned} \tag{10}$$

where $\mathbf{x}$ is the state vector of the signal, $u(n)$ is the process noise, $y(n)$ is the observation, $v(n)$ is the observation noise, $\mathbf{A}$ is the state transition matrix, and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix}, \tag{11}$$

$$\mathbf{b}^T = \mathbf{h} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}. \tag{12}$$

The Kalman filtering is first published in the 60s by Rudolf E. Kalman [58] [59] and since then has been extensively studied and applied in a large number of fields. The Kalman filter solutions can be found in many text books, e.g. [15]. For the fixed-interval smoothing problem, the Kalman theory also provides an interesting time-domain solution. Basically, Kalman smoothers first do a forward filtering followed by a backward filtering, and then combine the two filtering results. In this work, we use a "two-pass" Kalman smoothing algorithm which combines the last two steps in one sweep [60, p.572].

Although having been recognized as one of the major features, the non-stationary processing capability of the Kalman filter is, in many signal processing applications, not fully exploited. In speech processing, for example, the speech signals are known as highly non-stationary due to the fast movement of the articulators. The standard way of handling this non-stationarity is via short time processing. That is, to segment a

long sequence of speech signal into small frames, and assume local stationarity within each frame. As a consequence, the input process in the AR model is modeled as a stationary Gaussian process. As we have pointed out before, the impulse train structure in the input process is important to a good representation of the signal and should be modeled as either a non-Gaussian static process or a non-stationary Gaussian process. Thus we show, in paper C and D, that if the input process of voiced speech is modeled as a Gaussian process with rapidly varying variance, the Kalman filter (or smoother) can achieve a lower estimation MSE than the quasi-stationary Kalman filter. Different methods of estimating the slowly varying and fast varying parameters of the Kalman filter are also proposed in these two papers.

3.3 HMM filters and switching Kalman filters

The Hidden Markov Model (HMM) [61] [62] is a state-space model with discrete states. It is analogous to its continuous-state counterpart, the Kalman filter model in many ways. For example, both models use first-order Markovian dynamics to model state evolutions, and both observation processes are linear and Gaussian. The HMM can be expressed in a state-space form similar to the Kalman model (c.f. (10)), but with a nonlinear system equation:

$$\begin{aligned} x(n) &= f\big(x(n-1)\big) \\ y(n) &= x(n) + v(n), \end{aligned} \tag{13}$$

where $f(\cdot)$ is a nonlinear function. It is shown in [53] that the $f(\cdot)$ is a "winner-takes-all" nonlinearity, and that there is mapping between the representation using this nonlinearity and the one using a transition matrix. The HMM can also be seen as a Markovian-dynamical version of the Gaussian Mixture Model, which models non-Gaussianity with a sum of Gaussian pdfs. The HMM is widely used in modeling multi-mode systems with temporal structures in the transition of modes. The standard HMM filter estimates the discrete-valued Markov sequence hidden in white Gaussian noise. The filtering or smoothing is done with the forward-backward recursion [61].

Having the interesting capability of modeling the non-Gaussianity with a dynamical model, the HMM is ideal for modeling a non-Gaussian AR process with temporal structures in the input process. We designed a Hidden Markov-Autoregressive Model (HMARM), which cascades the HMM with the AR model, to model the temporal dependency in the input process and the dependency caused by the AR filter respectively. The motivation is that the conventional AR model only models correlation of the signal, which is a second order statistics, while the HMM can model higher order dependency that exists in the input process. The HMARM can also be seen as an extension of the HMM to explicitly model time correlation in the emitted samples. The conventional

HMM assumes that the emitted sample is independent of the previous ones. In the HMARM, the emitted samples are allowed to have correlation and the correlation is modeled by an AR(p) model. In this respect, a method in [63] provides an alternative of achieving a similar goal. In [63], the emission probability is modeled as a correlated multi-variate Gaussian pdf, which takes into account the correlation between the current sample and the previous one. This turns out to be a first order AR model.

The HMARM can be extended by introducing observation noise. We call it the Extended-HMARM (E-HMARM). When the signal is distorted by observation noise, the HMM filter alone is not sufficient, since it only deals with the process noise in the HMARM. An optimum nonlinear smoothing scheme is now needed. We propose to use a variant of the Switching Kalman filter with soft switching.

Switching Kalman filter is the collective name given to a group of methods (see [64] for a review). Conceptually, a switching Kalman filter models a system with a bank of linear models, and does optimum inference by switching between them or taking linear combinations of them. The switching decision is based on the probability of the hidden states that govern the linear models. Instead of switching all parameters of the system at every time instant as in [65] [66], or switching only the AR parameters frame-wise as in [67] [68], we switch the parameter of the input process at every time instant and keep the AR parameter constant within an analysis frame. In this way, the slowly varying AR parameters and the fast varying input process are modeled most efficiently (see Fig. 7). This is justified by our knowledge of many physical systems. For example, in the speech production system, the vocal tract (the filter) changes slowly compared to the movement of the vocal folds (the source); in communication systems, the physical channel (the filter) changes slowly compared to the transmitted symbols (the source).

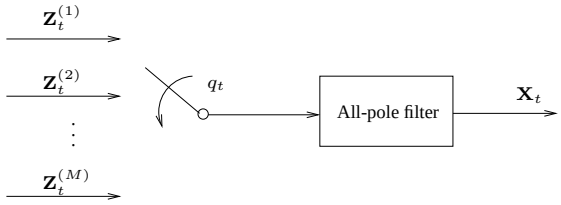


Figure 7: The switching AR signal model, where $\mathbf{Z}_t^M$ is the Mth constituent input process, and $\mathbf{X}_t$ is the observed signal. The state variable q_t controls the switch to select one input process at each time instant.

In paper E, F and G, we present the HMARM and E-HMARM models and algorithms for the filtering and system identification problems. Applications in speech analysis, noise robust spectra estimation, and blind channel equalization are demonstrated.

4 Parameter estimation

Parameter estimation, or system identification, is the process of learning the parameters of a system model given the observations and other information about the system. In the previous section we discussed the optimum filtering (or smoothing) problems for non-Gaussian AR signals, assuming known parameters. In practice, system parameters are generally unknown and need to be estimated before the signal can be estimated.

In the specific problem of AR model parameter estimation, the parameters can be grouped into two groups: the all-pole filter parameters and the excitation parameters. For a Gaussian AR model, the excitation process assumes an i.i.d. Gaussian pdf. Thus the only excitation parameter is the variance. For the Gaussian AR model, the filter parameter estimation problem and the excitation parameter estimation problem are decoupled. So all parameters can be estimated jointly. For non-Gaussian AR models, the excitation processes usually assume more complex models, and the filter parameters estimation problem and excitation parameters estimation problem are usually coupled. Most non-Gaussian AR model estimation algorithms estimate the two sets of parameters separately in iterative manners to reduce complexity [69] [70] [71]. In paper E and F, we show that the filter parameters and the excitation parameters can be jointly estimated by appropriately constraining the model.

In the following, we will review several major techniques for optimum estimation of parameters.

4.1 Least Squares methods

Least Squares (LS) is one of the most often used criterion in mathematical optimization. The LS method tries to find a set of parameters of the selected model that best fit to the measured data by minimizing the sum of the squares of the modeling error. It is shown by the Gauss-Markov theorem that the Least Squares estimator is the best linear unbiased estimator (BLUE) if the model is linear and if the modeling errors have zero mean and equal variance, and are uncorrelated. It is noteworthy that the LS criterion is a finite-sample approximate solution of the MSE criterion [4, p.91].

In the AR model parameter estimation problem, the optimum values for the parameters a_p are to be chosen such that the sum of the squared errors between the signal $x(t)$ and the predicted signal $\hat{x(t)}$ is minimized. The prediction here is a linear prediction using the previous p samples. Thus the cost function to be minimized is

$$C(\boldsymbol{\theta}) = \sum_{t=N_1}^{N_2} \Big[x(t) - \sum_{k=1}^{p} a_k x(t-k) \Big]^2 \tag{14}$$

where the $\boldsymbol{\theta} = [a_1, \cdots, a_p]^T$. The N_1 and N_2 are the indices of the boundary samples,

and the signal is assumed to be zeros outside of the boundaries. The vector $\boldsymbol{\theta}$ that minimizes the cost function can be shown to be

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^*\mathbf{X})^{-1}(\mathbf{X}^*\mathbf{x}) \tag{15}$$

where $\mathbf{x} = [x(N_1), \cdots, x(N_2)]^T$, and $\mathbf{X}$ is a Toeplitz matrix with $[0, x(N_1), \cdots, x(N_2)]^T$ as the first column. This result can also be obtained by writing the AR model in a matrix form:

$$\mathbf{x} = \mathbf{X}\boldsymbol{\theta} + \mathbf{u} \tag{16}$$

where the $\mathbf{x}$ and $\mathbf{X}$ are defined as same as above, and $\mathbf{u}$ is the vector of residuals. The residual is assumed to be a stationary process, and thus $\mathbf{u}$ can be seen as a perturbation vector. The parameter vector can be estimated by solving the perturbed linear system $\mathbf{x} \approx \mathbf{X}\boldsymbol{\theta}$ with the pseudo inverse, which results in (15).

There are two major variants that differ from each other by the choice of the boundaries: the autocorrelation method, which uses all available samples of the data frame in forming the $\mathbf{X}$, and the covariance method, which uses all samples except for the first p samples in forming the matrix $\mathbf{X}$. Notice that the matrix $\mathbf{X}^*\mathbf{X}$ is equivalent to the finite-sample estimate of the signal covariance matrix (up to a scaling factor).

The covariance method is found to be more accurate than the autocorrelation method when the data length is small [14]. The autocorrelation method though, is more popular in applications due to the existence of efficient implementation, e.g., the Levinson-Durbin algorithm (LDA) [72] [73]. An important observation here is that the autocorrelation method and the well known Yule-Walker method [74] lead to the same set of equations. For a Gaussian AR signal, the Yule-Walker method solves the optimum linear prediction problem by solving the Yule-Walker equations or normal equations:

$$\begin{bmatrix} r(0) & r(-1) & \cdots & r(-p) \\ r(1) & r(0) & & \vdots \\ \vdots & & \ddots & r(-1) \\ r(n) & \cdots & & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{17}$$

where $r(k)$ is the autocorrelation at lag k, and σ^2 is the variance of the input process. Due to the stationarity assumption, the autocorrelation matrix in the Yule-Walker system of equations is Toeplitz and Hermitian. The LDA exploits this structure and solve (3) in a recursive manner.

Both variants of the Least Squares method, as said, is based on the stationarity assumption. When applied to non-Gaussian or non-stationary signals, the bias and variance of the LS estimates are higher than that of the non-Gaussian estimators [17, p.147]. The cause of large bias and variance is the mismatch of Gaussian models to

non-Gaussian signal structures. For example, in the LPC analysis of voiced speech signals, the impulse train structure causes spectral sampling effects, which bias the estimated spectral envelope upwards at the harmonic frequencies and downwards at other frequencies. In paper E and F, a multi-state version of the Gaussian AR model has been developed, where the input process is modeled as several Gaussian processes controlled by a nonlinear switching mechanism. The resulting equation system is linear and can be seen as a multi-state version of the LS solution in (15).

Nonlinear Least Squares

The regression is called nonlinear regression when the regression model is not a linear function of the parameters. The method for nonlinear regression with the least squares criterion is called the Nonlinear Least Squares (NLS) method. The NLS method is often used in parameter estimation where the underlying nonlinear behavior of the process is well known. In general, solving the NLS problem requires numerical minimization techniques [75] such as Gauss-Newton method and grid searching.

The Multi-Pulse Linear Predictive Coding (MPLPC) is an example of the NLS method. The MPLPC is originally proposed by Atal and Remde [76] to optimally determine the impulse position and amplitude of the input process to the AR filter in the context of analysis-by-synthesis linear predictive coding. The criterion of the optimality is to minimize the sum of squares of modeling errors. Assuming that $h(n)$ is the (truncated) impulse response of the AR filter, and there are M pulses located at positions m_i with amplitudes g_i, $i \in [1, M]$, the cost function can be written as

$$C(g_i, m_i) = \sum_{t=1}^{N} \Big[x(t) - \sum_{i=1}^{M} g_i h(t - m_i) \Big]^2, \tag{18}$$

where N is the data frame length. Here the position parameter m_i is the nonlinear parameter. To solve the multi-dimensional nonlinear optimization (18) is difficult. A popular sub-optimal technique for this kind of problem is the Matching Pursuit (MP) technique, which decomposes the problem into a sequence of one-dimension optimizations. The MP finds the single best impulse, and subtract the effect of this impulse from the signal, and then find the next best impulse. Finding one impulse at a time is easy since it can be casted to a linear problem. Continuing until the required number of impulses are found, one gets a sequence of impulses that minimizes the cost function (18).

The MPLPC method is used in paper B and paper C for the estimation of temporal localization of power in the speech excitation. In using the MPLPC method for estimating the structure of the input process, the AR filter parameters need to be known or estimated first. The estimation of the AR parameters is done with the linear LS method

as introduced in the previous section. In paper B, the MPLPC model is modified such that the input process is a sum of a pulse train and a noise floor to better model the excitation of speech signals. The noise sequence and its amplitude are optimized as part of the nonlinear optimization.

The Total Least Squares method

In many practical problems, the output of the AR filter is distorted by observation noise. It is thus preferable to distinguish system noise and measurement noise since they are generated by different mechanisms. The ordinary LS method though, attributes all perturbations to the system noise. This can be seen clearly if the residual vector $\mathbf{u}$ in (16) can be written as a perturbation vector of $\mathbf{x}$:

$$\mathbf{x} + \boldsymbol{\Delta}\mathbf{x} = \mathbf{X}\boldsymbol{\theta}. \tag{19}$$

The Total Least Squares (TLS) is an extension to the LS method with an explicit perturbation to the signal matrix $\mathbf{X}$:

$$\mathbf{x} + \boldsymbol{\Delta}\mathbf{x} = (\mathbf{X} + \boldsymbol{\Delta}\mathbf{X})\boldsymbol{\theta}. \tag{20}$$

The TLS problem can be solved by first finding the $[\mathbf{X}; \mathbf{x}]$ that minimizes $[\boldsymbol{\Delta}\mathbf{X}; \boldsymbol{\Delta}\mathbf{x}]$ subject to $\mathbf{x} \in Range(\mathbf{X})$, and then solving for

$$\mathbf{x} = \mathbf{X}\boldsymbol{\theta}. \tag{21}$$

The minimization is usually done by finding the best lower rank approximation of the augmented matrix $[\mathbf{X} + \boldsymbol{\Delta}\mathbf{X}; \mathbf{x} + \boldsymbol{\Delta}\mathbf{x}]$, using the Singular Value Decomposition (SVD) technique.

It is shown in [77] (and the references therein) that the TLS estimator is a more robust parameter estimator than the LS estimator in noisy environments. Whereas, due to its very simple model, the TLS estimator can not utilize prior knowledge of the probability distributions of the system noise and measurement noise. If the Gaussian assumption is significantly violated, e.g., when outliers are present, the accuracy of the TLS deteriorates considerably and may be quite inferior to that of the LS estimates [77, p.5]. In this respect, the Bayesian analysis based on dynamical system models is a good alternative since it allows convenient modeling of system noise and measurement noise statistics.

4.2 Bayesian analysis of dynamic systems

One of the most popular dynamic model is the Kalman filter model, which is briefly reviewed in Section 3.2. Like the TLS, the Kalman filter model models both the system

noise and the measurement noise. But the Kalman filter model is more flexible in that the noise processes can be correlated, and non-stationary. More general dynamic models even allow non-Gaussian modeling of the noise, e.g., [78]. Some of the non-Gaussian MMSE estimation techniques mentioned in Section 1 have been or can be generalized to the dynamic models. Bayesian analysis though, is more used for signal estimation than parameter estimation, because the prior distribution of parameters are harder to learn than that of the signal waveforms. Thus the system identifications of Bayesian dynamic models are often treated as hidden data problems, and are solved via the EM algorithm. The principle is that, an MMSE estimator estimates the signal given the prior distributions of the system noise and the distribution of the measurement noise, and the parameters of the distributions of the noises are estimated by Maximum Likelihood estimators given the estimated clean signal. It can be shown that the iterations increase the likelihood function monotonically, so the resulting estimates of the parameters are equivalent to the ML estimates. The ML estimation and EM algorithm will be reviewed in the next section. Examples of identification of linear dynamic models can be found in [79] [53] [80]. In paper E and F, we derived blind system identification algorithms for non-Gaussian and nonlinear dynamic systems based on the EM paradigm.

4.3 The Maximum Likelihood method

The LS estimator reviewed in the previous section belongs to deterministic estimators since there is no statistics involved explicitly in its model. Introducing statistical models into the estimation is a way to improve estimation performance by exploiting statistical structure of the data. The Maximum Likelihood (ML) estimator is a popular statistical estimator for estimating parameters of an underlying probability distribution of a given data set.

In the ML estimation, the observation data $\mathbf{x}$ are assumed to be samples of a random process whose probability distribution are parameterized by a set of parameters $\boldsymbol{\theta}$. The ML estimator seeks the values of $\boldsymbol{\theta}$ that maximize the likelihood of the observations given the model. The likelihood is defined as

$$L(\boldsymbol{\theta}) \propto P(\mathbf{x}|\boldsymbol{\theta}). \tag{22}$$

The ML estimator is widely used in applications because it is easy to use and it is asymptoticly consistent and efficient. Asymptotic consistency and efficiency means that if the observation data length approaches infinity, the bias of the ML estimates approach zero and the variance approach the Cramer-Rao lower bound.

For the specific problem of ML estimation of Gaussian AR parameters, several works have been reported for the clean observation case [81] [82] [83] [84]. Even for Gaussian AR models, the exact ML estimators are nonlinear [84] [17], and are often

solved by numerical optimization or approximate ML estimations [17].

For the noisy observation case, the ML estimation of AR parameters are often done with iterative algorithms. A powerful iterative ML estimation technique called the Expectation-Maximization (EM) algorithm will be reviewed in the next section.

4.4 The Expectation-Maximization algorithm

The Expectation-Maximization (EM) algorithm is an iterative computation technique for maximum likelihood estimation. It is most suitable for incomplete data, or hidden data problems. Observation data corrupted by noise, or outputs of models whose latent variables are of real interest are examples of incomplete data. For an estimation problem that direct formulation of ML estimator is intractable or complicated, the problem can often be casted into a complete-data problem by appropriately choosing the complete data set, for which the ML estimation is more efficient. For example, while the maximization of the likelihood of the observation data need to be solved by computationally complex numerical optimizations, the maximization of the joint likelihood of the observation data and some other data can have a close form solution. The observation data and the extra data together are called the complete data. The extra data is usually unknown *a priori*, so the conditional mean (expectation) of the joint likelihood is maximized instead. Thus the EM algorithm iterates between the two steps, the maximization step (M-step) and the expectation step (E-step). The EM algorithm is shown to monotonically increase the likelihood at every iteration [85]. Thus it is an iterative ML estimator and enjoy the asymptotic property of the ML estimator.

Compared to other algorithms employing numerical optimization techniques such as gradient ascent methods and Newton type methods, the EM algorithm has the following advantages:

1. the EM algorithm has no such parameter as step size. Finding optimum time-dependent step size in the gradient ascent methods is a tricky and rather ad hoc process.

2. No need of finding Hessian and inverting Hessian as is needed in every iteration of the Newton type methods.

3. The EM algorithm is numerically stable with each iteration monotonically increasing the likelihood.

4. The E-step and M-step equations of an EM algorithm often give intuitive insights to the estimation problem, while the other numerical methods provide no such insight.

Generalized EM algorithms

In some problems, the M-step has no closed form solutions. In such cases, instead of choosing the parameters that maximize the expected likelihood of the complete data, the parameters can be chosen such that the expected likelihood is increased. It can be shown that this choice of parameters also increase the likelihood monotonically at each iteration [86, p.84]. This is called the Generalized EM (GEM) algorithm. One line of GEM algorithms use numerical maximization techniques in each M-step. Depending on the numerical methods used for the maximization, there exist different variants of GEM, such as the GEM Newton-Raphson algorithm [87] and the GEM gradient algorithm [88]. Another line of GEM uses the coordinate-ascent principle, which increases the multivariate likelihood function at each iteration by changing one parameter at a time [34]. If the free variable at each time is chosen to maximize the likelihood, the coordinate ascent converges to a local maximum [89].

The GEM algorithms, being easy to implement, have slower convergence rates than the exact EM algorithms, if exist. Also notice that in every iteration of the GEM, the expected likelihood is increased or locally maximized, unlike that in the exact EM the expected likelihood is globally maximized. So the GEM is more sensitive to the initial condition.

EM for parameter and signal estimation

In the application of EM algorithms to the estimation problem of noisy AR signals, the parameter estimation and signal estimation problems are integrated nicely in one theoretical framework. For Gaussian signal and noise, the complete data is usually defined as the concatenation of the observation and the clean signal. Using the signal model defined in (3) and (4), the complete data is denoted as

$$\mathbf{z} = \begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix}. \tag{23}$$

The parameters to be estimated, including the AR parameters $[a_1, \cdots, a_p]^T$, the process noise variance, and the measurement noise variance are denoted by the parameter vector $\boldsymbol{\theta}$.

In the M-step, the expected likelihood to be maximized is denoted by the Q-function

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(l)}) = E\{\log f(\mathbf{z}|\boldsymbol{\theta})|\mathbf{y}\}, \tag{24}$$

where $\boldsymbol{\theta}^{(l)}$ is the estimate of $\boldsymbol{\theta}$ at the l'th iteration, and the expectation is over the clean signal $\mathbf{x}$. The Q-function is maximized with respect to the parameter $\boldsymbol{\theta}$, resulting in a set of linear equations.

In the E-step, the expectation in (24), or the sufficient statistics of the signal, is calculated. This is usually done with the non-causal Wiener filter or the Kalman smoother.

At the stationary point of the algorithm, the ML estimates of all parameters and the MMSE estimates of the clean signal given the parameters are obtained.

Applying the EM algorithm to the estimation of Gaussian AR signals is first proposed by Feder, Oppenheim, and Weinstein [90] [91]. Though, a closely related iterative algorithm due to Lim and Oppenheim appears much earlier [92].

For non-Gaussian AR signals, the model for the excitation process is more complex, and either the M-step or the E-step can be nonlinear. For example, in [34] the non-Gaussian pdf is approximated by a mixture of Gaussian pdfs so that the filtering becomes a linear combination of linear filters, but the M-step requires solving a set of nonlinear equations. The solution in [34] is to use the generalized EM with coordinate ascent as described earlier.

Our approaches in paper E and F, are to impose further constraints on the excitation model. We show that when the mixture of Gaussian pdfs are constrained to have equal variance, the exact EM algorithm results in linear M-step and E-step. Further more, to exploit the temporal structure of the excitation process, we use an HMM to model the dynamics of the excitation process. It is shown that the EM identification algorithm for the HMM combined with the AR model has better convergence property and better estimation accuracy than the GMM ones for signals with temporal structure in the excitations.

Approximate EM algorithms

In the speech enhancement literature, there is a group of algorithms that have similar iterative structures to the EM algorithm. In [92] and [93], the algorithm iterates between the estimation of AR parameters and the estimation of the signal using Wiener filtering. In [94] and [95], The iterations are between AR parameter estimation and the Kalman filtering. In [96], a model for the long term correlation in the pitch is introduced. The parameters of the long term correlation model and the AR model are estimated from the noisy signal and then the Kalman filtering is done based on the estimated parameters. The algorithm iterates until convergence criterion is met. These algorithms are not designed explicitly based on the EM theory, but they are closely related to the EM algorithm and are conventionally seen as approximate EM algorithm.

In Paper D, we proposed an iterative algorithm based on Kalman filtering. Different from the above mentioned quasi-stationary EM methods, this method uses a Kalman filter model that has a non-stationary system noise with a rapidly varying variance. This method is an approximate EM algorithm. Another novelty is that the iteration is in a frame-wise sequential form. Instead of doing several iterations for each signal frame, the algorithm does the iterations along consecutive frames so that each frame is filtered

only once. The estimated spectrum of the previous frame is used in the initialization of the current frame estimation by a Weighted Power Spectral Subtraction (WPSS) initialization scheme. The WPSS filter combines the estimate of the previous frame with the current Power Spectral Subtraction estimate, much as the Decision-Directed method used in [40]. But it has different property than the Decision-Directed method because the signal phase is enhanced during the iteration due to the high resolution excitation modeling, while in the DD method phase is unprocessed. Due to the strong correlation between signal spectra of consecutive frames, the algorithm filters each frame only once and achieves the same gain as the conventional iterative scheme. In this way we can also obtain a good initialization for the iteration which is very important in iterative algorithms.

4.5 Higher Order Statistics based methods

Higher Order Statistics (HOS) based methods estimate model parameters using cumulants and their fourier transforms, known as polyspectra. HOS parameter estimation of LTI systems with non-Gaussian inputs has been extensively studied in the recent years. Works on AR estimation using HOS methods are found in [97] [38] [98] [37] [99]. In addition to the common properties of non-Gaussian processing techniques mentioned previously, major advantages of the HOS methods include:

1. The HOS based methods do not require a model for the pdf of the input process. Thus they are more general than methods assuming certain parametric forms for the distributions of input processes.

2. The HOS based methods are immune to Gaussian measurement noise. Either white or colored Gaussian noise can not degrade the estimation accuracy.

On the other hand, drawbacks associated with HOS methods are also significant:

1. HOS methods require longer data lengths than second-order method do. This is also a side effect of the non-parametric calculation of higher order statistics from samples. For many fast varying non-stationary signals, the calculation of high order cumulants are prohibitive in terms of data efficiency and computation efficiency.

2. HOS methods seldom use higher than 4th-order cumulants, because the higher the moment, the higher the estimator's variance will be [100] [38]. So they are unable to model nonlinearities higher than 4th-order.

In the speech processing literature, it is found that the higher order spectral analysis is associated with a higher spectral distortion compared to the second order ones [101]. This is due to the high variance of the HOS estimates given short frames of data. As

a principle, if any information/structure of the signal is known *a priori*, one should try to build it into a model, and then fit the model to the data. Good models help reduce estimation variance without need of long data.

5 Summary of contributions

The works included in this thesis are dedicated to solving the signal estimation and parameter estimation problems for non-Gaussian signals that posses rich temporal structures. We model such a signal as a stochastic process created by filtering a non-Gaussian input process with an all-pole filter. We term this model the generalized AR model since it resembles the standard AR model except that the input process can be of any probability distribution and can be temporally dependent. This model contains two parts: the all-pole filter with a moderate order models part of the temporal correlation of the signal, and a dynamical model is used to model the non-Gaussianity and correlation of the input process. Optimum non-Gaussian signal estimation and parameter estimation are addressed. A brief summary of our contributions on this subject is depicted in Fig. 8. Also shown in the diagram are the major established methods, and their positions in the big picture of AR signal estimation.

In papers A, B, C, and D, the focuses are on the optimum filtering of the non-Gaussian AR signals, based on extensions to the classical linear Gaussian filtering theories. We show that by treating the input process to the all-pole filter as a non-stationary process (i.e., dispensing with the quasi-stationarity assumption imposed on the input process), the temporal structures in the input process can be exploited for a better estimation of the signal. Thus by viewing a non-Gaussian process as a non-stationary Gaussian process, this approach solves the non-Gaussian signal estimation problem by modeling the non-stationarity. Specifically, the input process is modeled as a Gaussian process with zero-mean and a fast varying time-dependent variance. Parameters of the model are estimated before the filtering using the MPLPC technique, or using an iterative scheme, which iterates between parameter estimation and filtering.

In papers E, F and G, the non-Gaussianity of the input process is modeled by a GMM or an HMM model. The parameters of the GMM or the HMM, the all-pole filter parameters, and the measurement noise statistics are jointly estimated under the EM framework. The MMSE estimates of the non-Gaussian signal is obtained as a results of the E-step of the algorithm. The MMSE estimator we used here is a variant of the Switching Kalman Filter. The SKF is a nonlinear filter which combines a number of linear filters with a nonlinear switching function. When the GMM is used in the model, the non-Gaussianity in the input process is modeled without temporal dependency. When the HMM is used, the dynamics or the nonlinear temporal dependency in the input process is modeled. Thus it is possible in the HMARM model that the temporal de-

pendency in the input process and the temporal correlation caused by the all-pole filter are distinguished by the system identification algorithm. This is especially useful for de-convolution and equalization problems. Applications in speech analysis and channel equalization are demonstrated in the papers.

In summary, we propose, in this thesis, several non-Gaussian signal processing methods. These methods extend the classical linear Gaussian models in various ways to approach the non-Gaussian signal estimation problem with moderate additional complexities to the Gaussian ones by exploiting special signal structures. In these methods, the non-stationarity is fully exploited to model structures used to be modeled by non-Gaussianity and non-linearity.

References

[1] E. Mumolo and A. Carini, "Volterra adaptive prediction of speech with application to waveform coding," *European Transactions on Telecommunications*, vol. 6, No. 6, pp. 685–693, 1995.

[2] Y.-S. Zhang and D.-B. Li, "Volterra adaptive prediction of multipath fading channel," *Electronics Letters*, vol. 33, No. 9, pp. 754–755, 1997.

[3] S. Haykin, *Neural networks: a comprehensive foundation*. Englewood Cliffs, NJ: Macmillan Publishing Company, 1994.

[4] P. Stoica and R. Moses, *Introduction to spectral analysis*. Prentice Hall, 1997.

[5] J. Burg, "Maximum entropy spectral analysis," *PhD dissertation*, 1975.

[6] D. B. Percival and A. T. Walden, *Spectral analysis for physical applications: Multitaper and conventional univariate techniques*. Cambridge, UK: Cambridge University Press, 1993.

[7] G. P. Box and G. M. Jenkins, *Time series analysis - Forecasting and control*. San Francisco, CA: Holden-Day, 1976.

[8] T. W. Anderson, *The statistical analysis of time series*. New York: John Wiley & Sons, 1971.

[9] P. J. Brockwell and R. A. Davis, *Time series - theory and methods*. New York: Springer-Verlag, 1991.

[10] K. S. Shanmugan and A. M. Breipohl, *Random Signals - detection, estimation and data analysis*. John Wiley & Sons, Inc, 1988.

[11] L. Ljung, *System identification - Theory for the user*. Prentice Hall, Englewood Cliffs, N J, 1987.

[12] B. Atal and M. Schroeder, "Adaptive predictive coding of speech signals," *The Bell System Technical Journal*, pp. 1973–1987, Oct. 1971.

[13] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, pp. 561–580, Apr. 1995.

[14] S. L. Marple, *Digital spectral analysis with applications*. Englewood Cliffs, NJ: Prentice Hall, 1987.

[15] S. M. Kay, *Fundamentals of Statistical Signal Processing - Estimation Theory*. Prentice Hall PTR, 1993.

[16] K. K. Paliwal and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering," *Proc.of ICASSP 1987*, vol. 12, pp. 177–180, Apr. 1987.

[17] S. Kay and D. Sengupta, "Recent advances in non-gaussian autoregressive processes," in *Advances in spectrum analysis and array processing, vol.I*, S. Haykin, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[18] M. Grigoriu, *Applied non-Gaussian processes*. Englewood Cliffs, NJ: PTR Prentice Hall, 1976.

[19] G. K. Grunwald, R. J. Hyndman, L. Tedesco, and R. L. Tweedie, "Non-Gaussian conditional linear AR(1) models," *Australian & New Zealand Journal of Statistics*, vol. 42, issue 4, pp. 479–495, Dec. 2000.

[20] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, Inc., 2001.

[21] D. Sengupta and S. Kay, "Efficient estimation of parameters for non-Gaussian autoregressive processes," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37. No.6, pp. 785–794, 1989.

[22] C. Li and S. V. Andersen, "Blind identification of non-Gaussian Autoregressive models for efficient analysis of speech signals," *Proceedings of ICASSP*, 2006.

[23] C.-H. Lee, "Robust linear prediction for speech analysis," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 12, pp. 289–292, Apr. 1987.

[24] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1," *Vision Research*, vol. 37, No.23, pp. 3311–3325, 1997.

[25] Y. Li, A. Cichocki, and S.-I. Amari, "Analysis of sparse representation and blind source separation," *Neural Computation*, vol. 16, pp. 1193–1234, 2004.

[26] R. C. Reininger and J. D. Gibson, "Distribution of the two-dimensional DCT coefficients for images," *IEEE Trans. Commun.*, vol. COM-31, pp. 835–839, 1983.

[27] S. R. Kadaba, S. B. Gelfand, and R. L. Kashyap, "Recursive estimation of images using non-Gaussian autoregressive models," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 7. Issue 10, pp. 1439 – 1452, Oct. 1998.

[28] W. D. Penny and S. J. Roberts, "Variational Bayes for non-Gaussian autoregressive models," *IEEE intl. Workshop on Neural Networks for Signal Processing*, vol. 1, pp. 135–144, 2000.

[29] Y. Linde and R. M. Gray, "Fake process approach to data compression," *IEEE Trans. Commun.*, vol. COM-26, pp. 840–847, 1978.

[30] M. Shen and L. Sun, "The analysis and classification of phonocardiogram based on higher-order spectra," *IEEE Signal Processing Workshop on Higher-Oder Statistics*, pp. 29–33, 1997.

[31] S. Fang, W. Li, and D. Zhu, "Modeling and simulation of non-Gaussian correlated clutter," *Proc. of CIE Int. Conf. on Radar*, pp. 195–199, 1996.

[32] D. A. Hsu, "Long-tailed distributions for position errors in navigation," *Jour. Roy. Statist. Soc. ser. C*, vol. 28, pp. 62–72, 1979.

[33] G. Barnett, R. Kohn, and S. Sheather, "Bayesian estimation of an autoregressive model using Markov chain Monte Carlo," *Journal of Econometrics*, vol. 74(2), pp. 237–254, 1996.

[34] S. M. Verbout, J. M. Ooi, J. T. Ludwig, and A. V. Oppenheim, "Parameter estimation for autoregressive Gaussian-Mixture processes: the EMAX algorithm," *IEEE Trans. on Signal Processing*, vol. 46. No.10, pp. 2744–2756, 1998.

[35] M. S. Bartlett, *An introduction to stochastic processes*. London, UK: Cambridge University Press, 1955.

[36] D. R. Brillinger and M. Rosenblatt, "Computation and interpretation of kth-order spectra," in *Spectral analysis of time series*, B. Harris, Ed. New York: Wiley, 1967.

[37] J. M. Mendel, "Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications," *Proceedings of the IEEE*, vol. 79, No. 3, pp. 278–305, 1991.

[38] G. B. Giannakis and J. M. Mendel, "Identification of non-minimum phase systems using high-order statistics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 360–377, 1989.

[39] A. Gelb, *Applied optimal estimation*. Cambridge, Massachusetts: The M.I.T. Press, 1974.

[40] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[41] C. Breithaupt and R. Martin, "MMSE estimation of magnitude-squared DFT coefficients with supper Gaussian priors," *Proc.of ICASSP*, vol. 1, pp. 848–851, 2003.

[42] H. W. Sorensen and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, pp. 465–479, 1971.

[43] G. Kitagawa, "The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother," *Ann. Inst. Statist. Math.*, vol. 46, No.4, pp. 605–623, 1994.

[44] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proceedings*, vol. 140, No.2, pp. 107–113, Apr. 1993.

[45] S. J. Godsill and P. J. Rayner, "Robust noise reduction for speech and audio signals," *ICASSP Proc.*, pp. 625–628, 1996.

[46] E. R. Beadle and P. M. Djuric, "Parameter estimation for non-Gaussian autoregressive processes," *ICASSP Proc.*, vol. 5, pp. 3557–3560, Apr. 1997.

[47] R. Chen and J. S. Liu, "Mixture Kalman filters," *J. R. Statist. Soc. B*, vol. 62, Part 3, pp. 493–508, 2000.

[48] P. M. Djuric, J. H. Kotecha, F. Esteve, and E. Perret, "Sequential parameter estimation of time-varying non-Gaussian Autoregressive processes," *EURASIP Journal on Applied Signal Processing*, vol. 8, pp. 865–875, 2002.

[49] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-Interscience-IEEE, 1993.

[50] A. M. Kondoz, *Digital Speech, Coding for Low Bit Rate Communications Systems*. John Wiley & Sons, 1999.

[51] E. Denoel and J.-P. Solvay, "Linear Prediction of Speech with a Least Absolute Error Criterion," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, No. 6, pp. 1397–1403, 1985.

[52] M. Namba, H. Kamata, and Y. Ishida, "Neural Networks Learning with L1 Criteria and Its Efficiency in Linear Prediction of Speech Signals," *Proc. ICSLP '96*, vol. 2, pp. 1245–1248, 1996.

[53] S. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Neural Computation*, vol. 11. No.2, 1999.

[54] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, Inc, 1995.

[55] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation*. Prentice Hall, 2000.

[56] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, No. 2, pp. 113–120, Apr. 1979.

[57] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," *Proc. IEEE signal processing workshop on statistical signal processing*, pp. 496–499, Aug. 2001.

[58] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Engr. (ASME Trans.)*, vol. 82 D, pp. 35–45, 1960.

[59] ——, "New methods in Wiener filtering theory," *Proc. of the 1st. Symposium on Engineering Applications of Random Function and Probability*, pp. 270–388, 1963.

[60] G. Strang and K. Borre, *Linear Algebra, Geodesy and GPS*. Wellesley-Cambridge, U.S., 1997.

[61] L. R. Rabiner and B. H. Juang, "An introduction to Hidden Markov Model," *IEEE ASSP Magazine*, pp. 4–16, Jan. 1986.

[62] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, pp. 257–286, Feb. 1989.

[63] C. J. Wellekens, "Explicit time correlation in hidden Markov models for speech recognition," *ICASSP Proc.*, vol. 12, pp. 384–386, 1987.

[64] K. Murphy, "Switching Kalman filters," *Technical report, U. C. Berkeley*, 1998.

[65] J. B. Kim, K. Y. Lee, and C. W. Lee, "On the applications of the Interacting Multiple Model algorithm for enhancing noisy speech," *IEEE Trans. on Speech and Audio Processing*, vol. 8, No.3, pp. 493–508, May 2000.

[66] J. Deng, M. Bouchard, and T. Yeap, "Speech enhancement using a switching Kalman filter with a perceptual post-filter," *Proc. of ICASSP*, vol. I, pp. 1121–1124, 2005.

[67] B.-H. Juang and L. R. Rabiner, "Mixture autoregressive Hidden Markov Models for speech signals," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-33. No.6, pp. 1404–1413, 1985.

[68] A. Poritz, "Linear predictive hidden Markov models and the speech signal," *ICASSP'82*, vol. 7, pp. 1291–1294, 1982.

[69] S. Kay and D. Sengupta, "Statistically/computationally efficient estimation of non-Gaussian autoregressive processes," *ICASSP'87*, vol. 12, pp. 45–48, 1987.

[70] D. Burshtein, "Joint modeling and maximum-likelihood estimation of pitch and linear prediction coefficient parameters," *Journal of Acoustic Society of America*, vol. 91(3), pp. 1531–1537, Mar. 1992.

[71] Y. Zhao, X. Zhuang, and S.-J. Ting, "Gaussian mixture density modeling of non-Gaussian source for autoregressive process," *IEEE Trans. on Signal Processing*, vol. 43. No.4, pp. 894–903, 1995.

[72] N. Levinson, "The Wiener RMS (root mean square) criterion in filter design and prediction," *Journal of Math. and Physics*, vol. 25, pp. 261–278, 1947.

[73] J. Durbin, "The fitting of time series models," *Rev. Inst. Int. Stat.*, vol. 28, pp. 233–244, 1947.

[74] S. B. Kesler, *Modern spectral analysis II*. New York: IEEE Press, 1986.

[75] L. Ljung, "General structure of adaptive algorithms: adaptation and tracking," in *Adaptive system identification and signal processing algorithms*, N. Kalouptsidis and S. Theodoridis, Eds. UK: Prentice-Hall international, 1993.

[76] B. Atal and J. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," *Proc. of ICASSP 1982*, vol. 7, pp. 614–617, May 1982.

[77] S. V. Huffel and J. Vandewalle, *The total least squares problem: computational aspects and analysis*. Philadelphia: Society for Industrial and Applied Mathematics, 1991.

[78] A. Pole, M. West, and P. J. Harrison, "Nonnormal and nonlinear dynamic bayesian modeling," in *Bayesian analysis of time series and dynamic models*, J. C. Spall, Ed. New York: Marcel Dekker, Inc., 1988.

[79] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," *Technical Report CRG-TR-96-2, University of Toronto*, 1996.

[80] ——, "Variational Learning for Switching State-Space Models," *Neural Computation*, vol. 12. No.4, pp. 831–864, 2000.

[81] F. C. Schweppe, "Evaluation of likelihood functions for Gaussian signals," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 61–70, 1965.

[82] J. P. Burg, D. G. Luenburger, and D. L. Wenger, "Estimation of structured covariance matrices," *Proc. IEEE*, vol. 70, pp. 963–974, 1982.

[83] T. Kailath, B. Levy, L. Ljung, and M. Morf, "Fast time invariant implementation of Gaussian signal detectors," *IEEE Trans. Inform. Theory*, vol. IT-24, July 1978.

[84] L. T. McWhorter and L. L. Scharf, "Nonlinear Maximum Likelihood estimation of Autoregressive time series," *IEEE Trans. Signal Processing*, vol. 43, No.12, pp. 2909–2919, Dec. 1995.

[85] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc., Series B*, p. 138, 1977.

[86] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, Inc., 1997.

[87] S. N. Rai and D. E. Matthews, "Improving the EM algorithm," *Biometrics*, vol. 49, pp. 587–591, 1993.

[88] K. Lange, "A gradient algorithm locally equivalent to the EM algorithm," *Journal of the Royal Statistical Society*, vol. B, 57, pp. 425–437, 1995.

[89] D. G. Luenberger, *Linear and nonlinear programming, 2nd ed.* Addison Wesley, 1984.

[90] M. Feder, A. V. Oppenheim, and E. Weinstein, "Maximum likelihood noise cancellation using the EM algorithm," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 37, no.2, pp. 204–216, 1989.

[91] E. Weinstein, A. V. Oppenheim, and M. Feder, "Signal enhancement using single and multi-sensor measurements," *RLE Tech. Rep. 560, MIT, Cambridge, MA*, vol. 46, pp. 1–14, 1990.

[92] J. S. Lim and A. V. Oppenheim, "All-pole Modeling of Degraded Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASP-26, pp. 197–209, June 1978.

[93] J. H. L. Hansen and M. A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, 1991.

[94] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement," *IEEE Trans. on Signal Processing*, vol. 39, pp. 1732–1742, 1991.

[95] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. on Speech and Audio*, vol. 6, pp. 373–385, July 1998.

[96] Z. Goh, K. Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. on Speech and Audio Processing*, vol. 7, No.5, pp. 510–524, 1999.

[97] G. B. Giannakis, "On the identifiability of non-Gaussian ARMA models using cumulants," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 18–26, 1990.

[98] A. Swami and J. M. Mendel, "ARMA parameter estimation using only output cumulants," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1257–1265, 1990.

[99] K. K. Paliwal and M. M. Sondhi, "Recognition of noisy speech using cumulant-based linear prediction analysis," *Proc. ICASSP*, vol. 1, pp. 429–432, 1991.

[100] K. S. Lii and M. Rosenblatt, "A fourth-order deconvolution technique for non-gaussian linear processes," in *Multivariate Analysis-VI*, P. R. Krishnaiah, Ed. New York: Elsevier Science, 1985, pp. 395–410.

[101] J. M. Salavedra, E. Masgrau, A. Moreno, J. Estarellas, and X. Jove, "Robust coefficients of a higher order AR modeling in a speech enhancement system using parameterized Wiener filtering," *Proc. 7th Mediterranean Electrotechnical Conference*, vol. 1, pp. 69–72, 1994.

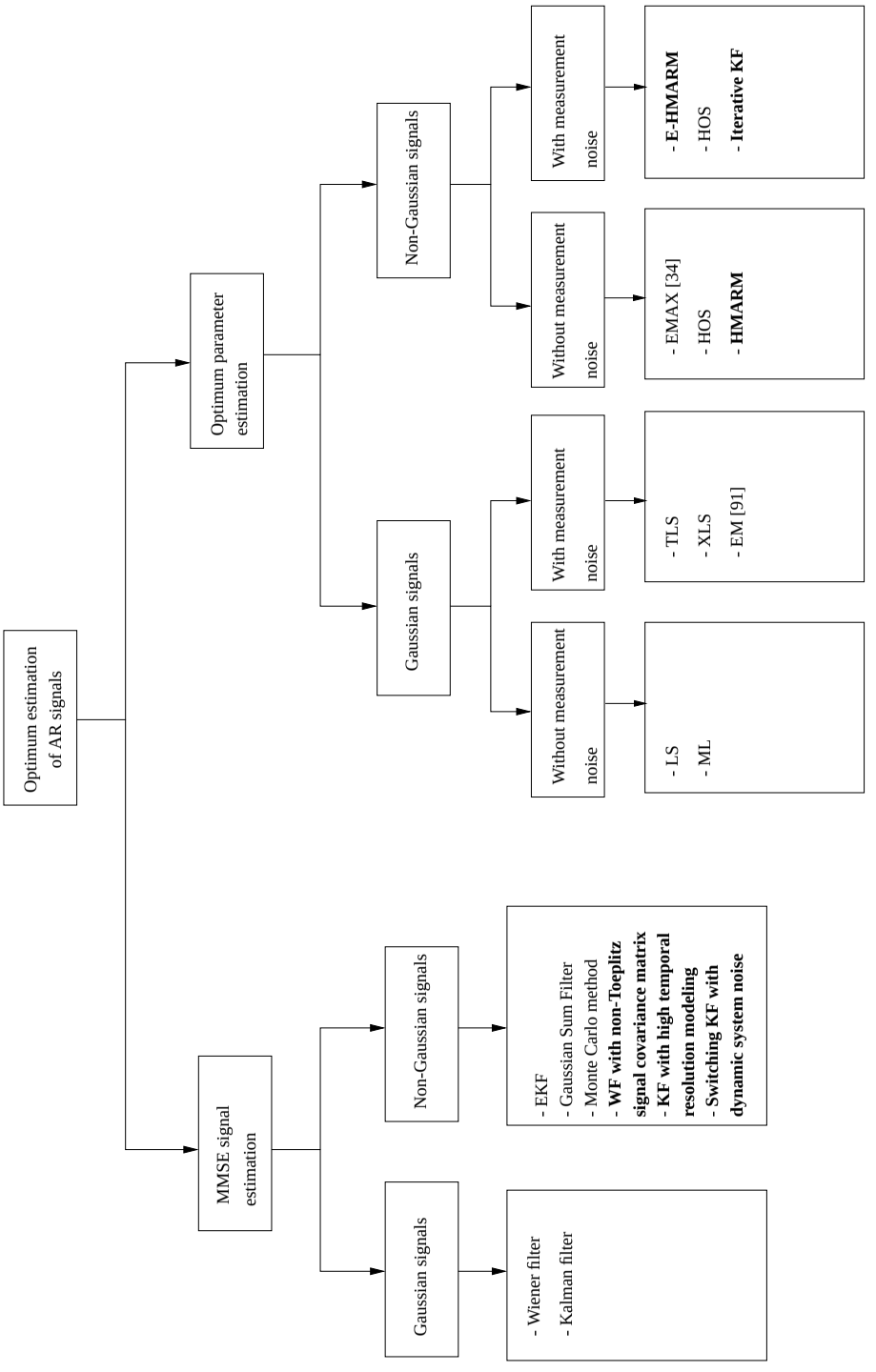


Figure 8: A brief summary of methods for AR signal estimation, and contributions of this work (in boldface).

Part II

Papers

Paper A

Inter-frequency Dependency in MMSE Speech Enhancement

Chunjian Li and Søren Vang Andersen

The paper has been published in
Proceedings of the 6th Nordic Signal Processing Symposium, pp. 200-203.
June 9-11, 2004. Espoo, Finland.

The layout has been revised.

Abstract

In this paper an MMSE estimator of the complex short-time spectrum is considered for optimum noise reduction of speech. The correlation between frequency components is exploited to improve the estimation, especially of those components with low local SNR. Furthermore, by making use of both spectral envelope and time envelope, the estimator is able to suppress noise power in frequency domain and time domain simultaneously. The performance of the resulting estimator is found to be superior to the non-causal IIR Wiener filter. The enhanced signal suffers less spectral distortion, while achieving a lower mean squared error than the Wiener filter.

1 Introduction

In recent years, several MMSE approaches to speech enhancement appeared, including the non-causal IIR Wiener filter [1], the MMSE STSA estimator [2], and MMSE estimator using non-Gaussian priors [3]. Most of them can be characterized as short-time spectral amplitude estimators. A common characteristic of these methods is that they only process the spectral amplitude and use the noisy phase spectra to generate the enhanced signals (except for [3], in which the real parts and imaginary parts of the DFT coefficients are independently estimated). As an example, take the non-causal IIR Wiener filter with transfer function defined by

$$H_{WF}(\omega) = \frac{P_{ss}(\omega)}{P_{ss}(\omega) + P_{vv}(\omega)} \tag{1}$$

where $P_{ss}(\omega)$ and $P_{vv}(\omega)$ denote the power spectral density of the speech signal and the uncorrelated additive noise, respectively. Hereafter we refer to (1) as the Wiener filter or WF. The transfer function of the WF is of zero phase and therefore it leaves the phase unprocessed. In addition, the WF does not exploit any inter-frequency dependency. This is a consequence of the stationarity assumption, and is another common point of the established MMSE approaches. One reason for not processing the phase spectrum is that phase is found to play a less important role in the human perception of speech [4]. An approximate threshold of phase perception was found in [4] corresponding to a local SNR of about 6 dB. If a frequency component in a frame has a local SNR higher than 6 dB, the phase distortion is not audible. The second common point comes as a consequence of assuming the speech frame to be infinitely long and stationary [5]. Although speech signals are known to be non-stationary and short-time processing is applied, this assumption is widely used in order to simplify the estimator.

In this paper we show that if these two restrictions are removed, better estimators are obtained.

2 Phase spectrum and inter-frequency dependency

The motivation for involving phase information in the MMSE estimator is that, first of all, phase distortion is audible with low SNR speech. Processing low SNR speech with an estimator working only on the spectral amplitude brings reverberant effect and roughness to the enhanced speech. Recent works [6, 7] confirm that, especially for the voiced male speech, phase information is of clear perceptual importance. Moreover, the phase noise causes amplitude spectrum distortion through phase modulation when the signals are short-time processed using the overlap-add method. The rise of the spectrum in the valley between pitch harmonics causes audible artifacts and higher residual noise.

Secondly, phase coherence in the voiced speech is a significant source of correlation between frequency components. Two sources of correlation among frequency components can be identified. One is the finite-length window effect. It is known that the infinite Fourier matrix is the eigenvector matrix of an infinite Toeplitz matrix [8]. If we denote the covariance matrix of the speech samples, the inverse Fourier matrix, and the covariance matrix of the frequency components as $\mathbf{C_s}$, $\mathbf{F}$, and $\mathbf{C}_\theta$, respectively, we can write the covariance matrix as $\mathbf{C}_\theta = \mathbf{F}\mathbf{C_s}\mathbf{F^H}$. When $\mathbf{C_s}$ is a Toeplitz matrix, if the frame length of the Fourier analysis approaches infinity, $\mathbf{C}_\theta$ will become diagonal. However in general the speech signal is non-stationary, and very long windows are not applicable. The finite-length window effect causes the covariance matrix $\mathbf{C}_\theta$ to be generally non-diagonal. Therefore correlation exist among the frequency components. The second, and more interesting source of correlation is the phase coherence in voiced speech. Voiced speech can be modeled as an excitation pulse train filtered by an all-pole filter. The phase of the pulse train is approximately linear at pitch harmonic frequencies. After the filtering, the coherence in phase is maintained to some extend. If the phase coherence is lost, the voiced speech sounds reverberant [9]. The coherence in phase corresponds to energy localization in the time domain, which can be modeled by a time envelope.

Because of the importance of phase stated above, and because the optimum amplitude estimator and the optimum phase estimator do not coexist [2], we formulate the MMSE estimator as an estimate of the complex Fourier coefficients instead of independently derived spectral amplitude and phase estimators as in [2] or independent real parts and imaginary parts as in [3].

3 MMSE estimator with time and frequency envelopes

The key feature of the new MMSE estimator is modeling the covariance matrix $\mathbf{C}_\theta$ as a full matrix instead of a diagonal matrix as in the WF. We will show the frequency domain MMSE estimator first and then transform it to time domain.

We use the following statistical model and problem formulation. The DFT coefficients of each speech segment are modeled as complex Gaussian random variables with zero mean and varying variance. Let $y(n,k)$, $s(n,k)$, $v(n,k)$ denote the n'th sample of noisy observation, speech, and additive white Gaussian noise of the k'th frame, respectively. Then

$$y(n,k) = s(n,k) + v(n,k). \tag{2}$$

Let $\theta(m,k)$ represent the m'th DFT coefficient of the k'th frame, defined by $\theta(m,k) = \sum_{n=0}^{N} s(n,k)exp(-j2\pi nm/N)$. For compactness we use vector representation and omit the index in the following discussion. Let $\mathbf{y}$, $\boldsymbol{\theta}$, $\mathbf{v}$, and $\mathbf{F}$ denote the vectors of y, θ, v and the inverse Fourier matrix respectively. Then (2) can be written as

$$\mathbf{y} = \mathbf{F}\boldsymbol{\theta} + \mathbf{v}. \tag{3}$$

The MMSE estimator can be shown to be the conditional mean [10]

$$\begin{aligned} \hat{\theta} &= E(\boldsymbol{\theta}|\mathbf{y}) \\ &= \mathbf{C}_\theta \mathbf{F}^{\mathbf{H}} (\mathbf{F}\mathbf{C}_\theta \mathbf{F}^{\mathbf{H}} + \mathbf{C}_{\mathbf{v}})^{-\mathbf{1}} \mathbf{y} \end{aligned} \tag{4}$$

where $(\cdot)^H$ denotes the Hermitian transpose and $\mathbf{C_v}$ denotes the covariance matrix of the noise $\mathbf{v}$. The covariance matrix $\mathbf{C}_\theta$ is generally unknown and must be replaced with an estimate. We propose here an approach to the estimation of $\mathbf{C}_\theta$ from the all-pole model of the speech. Let $q/A(z)$ denote the transfer function of the all pole model. Let $\mathbf{H}$ be the corresponding synthesis filter matrix derived from the all-pole model, and $\mathbf{r}$ be the residual vector, such that

$$\mathbf{s} = \mathbf{H}\mathbf{r}. \tag{5}$$

Since the residual is a white noise sequence with unit variance (for voiced speech it is a few impulses present periodically in the white noise), the covariance matrix $\mathbf{C_r}$ of $\mathbf{r}$ can be written as a diagonal matrix with the squared residual as the diagonal elements [1]. Once $\mathbf{C_r}$ is obtained, $\mathbf{C_s}$ and $\mathbf{C}_\theta$ can easily be found. We have

$$\mathbf{C_s} = \mathbf{H}\mathbf{C_r}\mathbf{H^H} \tag{6}$$

$$\mathbf{C}_\theta = \mathbf{F^H}\mathbf{C_s}\mathbf{F}. \tag{7}$$

Inserting (7) in (4) gives the MMSE short-time spectral estimator.

Fig.1 shows how the covariance matrix $\mathbf{C}_\theta$ estimated by this approach differs from the diagonal matrix underlying the standard WF. We can see that the off-diagonal elements are generally non-zero. At the brims of the matrix the cross-correlations are

[1] Here we ignore the long term correlation of the residual.

significant. This represents the windowing effect caused by the high spectral power at low frequencies. More interestingly, we see how inter-frequency dependency, especially between neighboring formants show up as significant off-diagonal elements in the covariance matrix. It is well known that a properly chosen window can reduce the correlation between frequency components but can not eliminate it. In Fig.1 a Hanning window is used, and we see that the remaining correlation is still significant and can be exploited to improve the estimator.

The frequency domain MMSE estimator given by (4) is mainly for the purpose of demonstrating the difference to the WF made by a full covariance matrix. In the estimation of the speech waveform, (4) is transformed back to time domain, giving the desired time domain MMSE estimator,

$$\hat{\mathbf{s}} = \mathbf{C_s}(\mathbf{C_s} + \mathbf{C_v})^{-\mathbf{1}}\mathbf{y}. \tag{8}$$

Estimating the diagonal elements of $\mathbf{C_r}$ is equivalent to estimating the residual power distribution over the time axis. It can also be seen as estimating phase from the residual, because the power spectrum of the residual is known to be white. Estimating the squared residual from noisy observation is difficult. Our solution is to estimate the time envelope of the squared residual with simple shapes, i.e. a constant floor plus some pulses located periodically. These varying variances of residual represent time localization of energy. This is a major difference to the WF, which can be seen as using constant residual variance because of the stationary assumption. We estimate the residual envelope in a simple but effective way. The noisy speech signal is first lowpass filtered with cut-off frequency of 800 Hz. A 3-tap whitening filter is found by applying linear prediction on the filtered signal. The output of the low pass filter is then filtered by the whitening filter to get a reference residual. The position of the maximum in the reference residual is chosen as the first impulse position of the estimated residual envelope. According to an estimate of the pitch period the positions of remaining impulses are found. A pre-defined pulse shape is put on every impulse position. The pulse shape is chosen to be wider and smoother than a true residual impulse in order to gain robustness against error in estimating the impulse positions. The rest of the residual will be approximated with a constant whose amplitude is decided by keeping the average power of the estimated residual equal to unity. The estimation of the residual envelope is only needed for voiced frames. Fig.2 shows an example of the estimated residual envelope.

Because the above described MMSE estimator requires a spectral envelope and a temporal envelope as the prior knowledge, we hereafter refer to it as the Time-Frequency Envelope MMSE (TFE-MMSE) estimator.

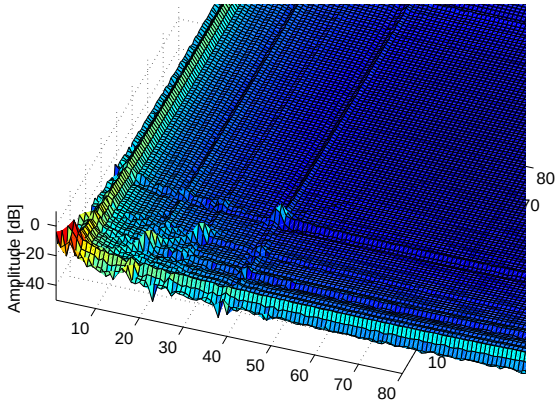


Figure 1: Amplitude plot of the covariance matrix $\mathbf{C}_\theta$. Matrix size is 160 by 160 (only one quarter of the matrix is shown).

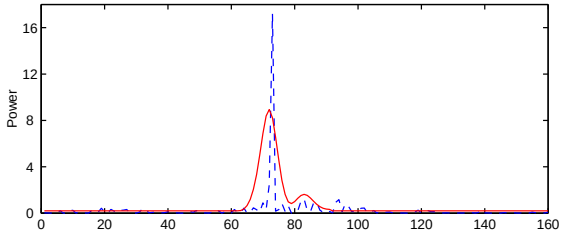


Figure 2: The squared residual (dashed) and the estimated envelope (solid).

4 results

We first compare the performance of the TFE-MMSE estimator and the WF based on known spectral envelope of the signal. Since the purpose is to show that using the extra information about phase (or energy localization in time) it is possible to achieve lower mean squared error and lower spectral distortion at the same time, we first use known spectral envelopes for both estimators.

Both estimators run with 30 sentences from different speakers (15 male and 15 female) from the TIMIT database added with artificial white Gaussian noise at a signal-to-noise ratio of 0 dB. All sentences are 16kHz sampled, and segmented into frames of 160 samples. For the TFE-MMSE estimator, the time envelopes of the residual are estimated from noisy observations using the method described in section 3. For the output of both estimators, the SNR, Segmental SNR (segSNR) and Log-Spectral Distortion (LSD) to the original signal spectrum are calculated. The SNR is defined as the ratio of

the total signal power to the total noise power in the sentence. The segSNR is defined as the average ratio of signal power to noise power per frame, omitting frames with a power more than 30 dB below average power. The LSD is defined as the distance between two log-scaled DFT spectra summed over all frequencies. The LSD is calculated only for voiced frames since for the unvoiced frames both estimators are identical.

From Table 1 we see consistent improvement of the TFE-MMSE estimator over WF in all three measurements. Fig.3 shows the signal spectrum of a voiced frame comparing with the spectrum of the output of the two estimators. Only the lower frequency half is plotted to show the details of the harmonic structure. It is seen that the TFE-MMSE estimator preserves the harmonic structure better than the WF.

To verify the performance in a practical scenario, estimated LPC coefficients are also used in the comparison. The LPC coefficients are estimated by a method similar to the decision directed method in [2]. The experimental setup is identical to the above one, except that input SNR is now set to 10 dB. Table 2 shows the results. Significant improvements are observed with the segSNR measurement. The LSD of the TFE-MMSE estimator also improves significantly over the WF. Informal listening experiments show that the reduction of spectral distortion is significant.

	Male			**Female**		
	SNR	segSNR	LSD	SNR	segSNR	LSD
WF	10.73	5.21	290	10.57	5.59	347
TFE-MMSE	11.24	5.48	265	10.85	5.71	315
Improv.	0.51	0.27	25	0.28	0.12	32

Table 1: Performance of WF and the TFE-MMSE estimator with known AR coefficients. All SNR measures are in dB. Input SNR is 0 dB. Results are averaged over 30 sentences (by 15 male and 15 female speakers).

	Male			**Female**		
	SNR	segSNR	LSD	SNR	segSNR	LSD
WF	15.65	8.73	245	15.38	9.30	303
TFE-MMSE	16.71	9.42	183	16.48	9.83	231
Improv.	1.06	0.70	62	1.10	0.53	72

Table 2: Performance of WF and the TFE-MMSE estimator with estimated AR coefficients. Input SNR is 10 dB. Results are averaged over 30 sentences (by 15 male and 15 female speakers).

5 Discussion

In the first part of this paper we stated the motivation of formulating an MMSE joint estimator of amplitude and phase spectrum, i.e., phase is of perceptual importance for low SNR sources, and estimating phase provides the additional information about the correlation of DFT coefficients which improves the amplitude spectrum estimation in return.

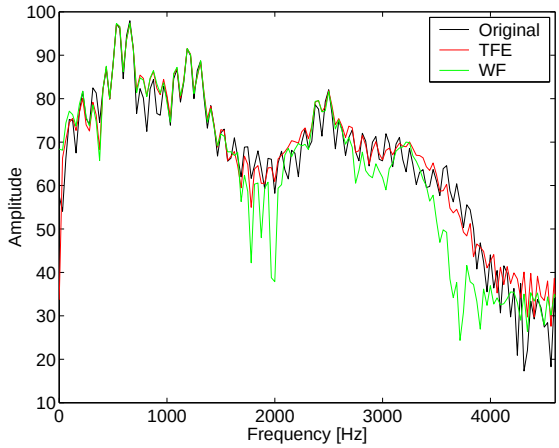


Figure 3: A comparison of amplitude spectrum for the output of WF and the TFE-MMSE estimator to the original signal spectrum.

We have avoided the widely used assumption of independent frequency components. This is justified by the fact that both finite-length window effect and time localization of energy (caused by phase coherence) in the voiced speech introduce correlation among the frequency components. Phase is known as hard to estimate, so we re-formulate the problem into estimating time envelope of the residual power. The MMSE joint spectral estimator (4) shows us that a full covariance matrix can exploit the inter-frequency dependency, achieving a better spectrum estimate. The algorithm is finally implemented as a time domain MMSE estimator (8).

The performance of the TFE-MMSE estimator and Wiener filter are compared based on known LPC coefficients as well as estimated ones. The TFE-MMSE estimator shows higher SNR and less spectral distortion than the WF. In the case of using estimated LPC coefficients, the improvement of segmental SNR and spectral distortion of the TFE-MMSE estimator over the WF is even more significant. This is because the spectral suppression and the temporal suppression benefit from each other making a better joint estimator.

References

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[3] R. Martin, “Speech Enhancement Using MMSE Short Time Spectral Estimation With Gamma Distributed Speech Priors,” *Proc.of ICASSP 2002*, vol. 1, pp. 253–256, May 2002.

[4] P. Vary, “Noise Suppression By Spectral Magnitude Estimation - Mechanism and Theoretical Limits,” *Signal Processing 8*, pp. 387–400, May 1985.

[5] W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*. New York: McGraw-Hill, 1958.

[6] H. Pobloth and W. B. Kleijn, “On Phase Perception in Speech,” *Proc.of ICASSP 1999*, vol. 1, pp. 29–32, Mar. 1999.

[7] J. Skoglund, W. B. Kleijn, and P. Hedelin, “Audibility of Pitch-Synchronously Modulated Noise,” *Speech Coding For Telecommunications Proceeding, IEEE*, vol. 7-10, pp. 51–52, Sept. 1997.

[8] R. M. Gray, “Toeplitz and Circulant Matrices: A review,” *Foundations and Trends in Communications and Information Theory*, vol. 2, Issue 3, pp. 155–239, 2006.

[9] T. F. Quatieri and R. J. McAulay, “Phase Coherence in Speech Reconstruction for Enhancement and Coding Applications,” *Proc.of ICASSP 1989.*, vol. 1, pp. 207–210, May 1989.

[10] S. M. Kay, *Fundamentals of Statistical Signal Processing - Estimation Theory*. Prentice Hall PTR, 1993.

Paper B

A Block Based Linear MMSE Noise Reduction with a High Temporal Resolution Modeling of the Speech Excitation

Chunjian Li and Søren Vang Andersen

The paper has been published in
EURASIP Journal on Applied Signal Processing, Special Issue on DSP in Hearing Aids and Cochlear Implants, vol. 2005:18, pp. 2965-2978, October 2005.

The layout has been revised.

1 Abstract

A comprehensive Linear Minimum Mean Squared Error (LMMSE) approach for parametric speech enhancement is developed. The proposed algorithms aim at joint LMMSE estimation of signal power spectra and phase spectra, as well as exploitation of correlation between spectral components. The major cause of this inter-frequency correlation is shown to be the prominent temporal power localization in the excitation of voiced speech. LMMSE estimators in time domain and frequency domain are first formulated. To obtain the joint estimator, we model the spectral signal covariance matrix as a full covariance matrix instead of a diagonal covariance matrix as is the case in the Wiener filter derived under the quasi-stationarity assumption. To accomplish this, we decompose the signal covariance matrix into a synthesis filter matrix and an excitation matrix. The synthesis filter matrix is built from estimates of the all-pole model coefficients, and the excitation matrix is built from estimates of the instantaneous power of the excitation sequence. A decision-directed Power Spectral Subtraction method and a modified Multi-Pulse Linear Predictive Coding (MPLPC) method are used in these estimations, respectively. The spectral domain formulation of the LMMSE estimator reveals important insight about inter-frequency correlations. This is exploited to significantly reduce computational complexity of the estimator. For resource-limited applications such as hearing aids, the performance-to-complexity tradeoff can be conveniently adjusted by tuning the number of spectral components to be included in the estimate of each component. Experiments show that the proposed algorithm is able to reduce more noise than a number of other approaches selected from the state-of-the-art. The proposed algorithm improves the segmental SNR of the noisy signal by 13 dB for the white noise case with an input SNR of 0 dB.

2 Introduction

Noise reduction is becoming an important function in hearing aids in recent years thanks to the application of powerful DSP hardware and the progress of noise reduction algorithm design. Noise reduction algorithms with high performance-to-complexity ratio have been the subject of extensive research study for many years. Among many different approaches, two classes of single-channel speech enhancement methods have attracted significant attention in recent years because of their better performance compared to the classic spectral subtraction methods (A comprehensive study of Spectral Subtraction methods can be found in [1]). These two classes are the frequency domain block based Minimum Mean Squared Error (MMSE) approach and the signal subspace approach. The frequency domain MMSE approach includes the non-causal IIR Wiener filter [2], the MMSE Short-Time Spectral Amplitude (MMSE-STSA) estimator [3], the MMSE Log-Spectral Amplitude (MMSE-LSA) estimator [4], the Constrained Iterative Wiener

Filtering (CIWF) [5], and the MMSE estimator using non-Gaussian priors [6]. These MMSE algorithms all rely on an assumption of quasi-stationarity and an assumption of uncorrelated spectral components in the signal. The quasi-stationarity assumption requires short time processing. At the same time, the assumption of uncorrelated spectral components can be warranted by assuming the signal to be infinitely long and wide-sense stationary [7] [8]. This infinite data length assumption is in principle violated when using the short-time processing, although the effect of this violation may be minor (and is not the major issue this paper addresses). More importantly, the wide-sense stationarity assumption within a short frame does not well model the prominent temporal power localization in the excitation source of voiced speech due to the impulse train structure. This temporal power localization within a short frame can be modeled as a non-stationarity of the signal that is not resolved by the short-time processing. In [9], we show how voiced speech is advantageously modeled as non-stationary even within a short frame, and that this model implies significant inter-frequency correlations. As a consequence of the stationarity and long frame assumptions, the MMSE approaches model the frequency domain signal covariance matrix as a diagonal matrix.

Another class of speech enhancement methods, the signal subspace approach, implicitly exploits part of the inter-frequency correlation by allowing the frequency domain signal covariance matrix to be non-diagonal. This class includes the Time Domain Constraint (TDC) linear estimator and Spectral Domain Constraint (SDC) linear estimator [10], and the Truncated Singular Value Decomposition (TSVD) estimator [11]. In [10], the TDC estimator is shown to be an LMMSE estimator with adjustable input noise level. When the TDC filtering matrix is transformed to the frequency domain, it is in general non-diagonal. Nevertheless, the known signal subspace based methods still assume stationarity within a short frame. This can be seen as follows. In TDC and SDC the noisy signal covariance matrices are estimated by time averaging of the outer product of the signal vector, which requires stationarity within the interval of averaging. The TSVD method applies singular value decomposition to the signal matrix instead. This can be shown to be equivalent to the eigen decomposition of the time averaged outer product of signal vectors. Compared to the mentioned frequency domain MMSE approaches, the known signal subspace methods implicitly avoid the infinite data length assumption, so that the inter-frequency correlation caused by the finite length effect is accommodated. However, the more important cause of inter-frequency correlation, i.e., the non-stationarity within a frame is not modeled.

In terms of exploiting the masking property of the human auditory system, the above mentioned frequency domain MMSE algorithms and signal subspace based algorithms can be seen as spectral masking methods without explicit modeling of masking thresholds. To see this, observe that the MMSE approaches shape the residual noise (the remaining background noise) power spectrum to one more similar to the speech power spectrum, thereby facilitating a certain degree of masking of the noise. In general, the

MMSE approaches attenuate more in the spectral valleys than the spectral subtraction methods do. Perceptually, this is beneficial for high pitch voiced speech, which has sparsely located spectral peaks that are not able to mask the spectral valley sufficiently. The signal subspace methods in [10] are designed to shape the residual noise power spectrum for a better spectral masking, where the masking threshold is found experimentally. Auditory masking techniques have received increasing attention in recent research of speech enhancement [12–14]. While the majority of these works focus on spectral domain masking, the work in [15] shows the importance of the temporal masking property in connection with the excitation source of voiced speech. It is shown that noise between the excitation impulses is more perceivable than noise close to the impulses, and this is especially so for the low pitch speech for which the excitation impulses locates temporally sparsely. This temporal masking property is not employed by current frequency domain MMSE estimators and the signal subspace approaches.

In this paper, we develop an LMMSE estimator with a high temporal resolution modeling of the excitation of voiced speech, aiming for modeling a certain non-stationarity of the speech within a short frame, which is not modeled by quasi-stationarity based algorithms. The excitation of voiced speech exhibits prominent temporal power localization, which appears as an impulse train superimposed with a low level noise floor. We model this temporal power localization as a non-stationarity. This non-stationarity causes significant inter-frequency correlation. Our LMMSE estimator therefore avoids the assumption of uncorrelated spectral components, and is able to exploit the inter-frequency correlation. Both the frequency domain signal covariance matrix and filtering matrix are estimated as complex-valued full matrices, which means that the information about inter-frequency correlation are not lost and the amplitude and phase spectra are estimated jointly. Specifically, we make use of the linear prediction based source-filter model to estimate the signal covariance matrix, upon which a time domain or frequency domain LMMSE estimator is built. In the estimation of the signal covariance matrix, this matrix is decomposed into a synthesis filter matrix and an excitation matrix. The synthesis filter matrix is estimated by a smoothed power spectral subtraction method followed by an autocorrelation Linear Predictive Coding (LPC) method. The excitation matrix is a diagonal matrix with the instantaneous power of the LPC residual as its diagonal elements. The instantaneous power of the LPC residual is estimated by a modified Multi-Pulse Linear Predictive Coding (MPLPC) method. Having estimated the signal covariance matrix, we use it in a vector LMMSE estimator. We show that by doing the LMMSE estimation in the frequency domain instead of in time domain, the computational complexity can be reduced significantly due to the fact that the signal is less correlated in the frequency domain than in the time domain. Compared to several quasi-stationarity based estimators, the proposed LMMSE estimator results in a lower spectral distortion to the enhanced speech signal while having higher noise reduction capability. The algorithm applies more attenuation in the valleys between pitch impulses in time

domain, while small attenuation is applied around the pitch impulses. This arrangement exploits the temporal masking effect, and results in a better preservation of abrupt rise of the waveform amplitude while maintaining a large amount of noise reduction.

The rest of this paper is organized as follows. In Section 3 the notations and assumptions used in the derivation of LMMSE estimators are outlined. In Section 4, the non-stationary modeling of the signal covariance matrices is described. The algorithm is summarized in Section 5. In Section 6, the computational complexity of the algorithm is reduced by identifying an interval of significant correlation and by simplifying the modified MPLPC procedure. Experimental settings, objective, and subjective results are given in Section 7. Finally, Section 8 discusses the obtained results.

3 Background

In this section, notations and statistic assumptions for the derivation of LMMSE estimators in time and frequency domain are outlined.

3.1 Time domain LMMSE estimator

Let $y(n,k)$, $s(n,k)$, $v(n,k)$ denote the n'th sample of noisy observation, speech, and additive noise (uncorrelated with the speech signal) of the k'th frame, respectively. Then

$$y(n,k) = s(n,k) + v(n,k).$$

Alternatively, in vector form we have

$$\mathbf{y} = \mathbf{s} + \mathbf{v}, \tag{1}$$

where boldface letters represent vectors and the frame indices are omitted to allow a compact notation. For example $\mathbf{y} = [y(1,k), y(2,k), \cdots, y(N,k)]^T$ is the noisy signal vector of the k'th frame, where N is the number of samples per frame.

To obtain linear MMSE estimators, we assume zero mean Gaussian PDF's for the noise and the speech processes. Under this statistic model the LMMSE estimate of the signal is the conditional mean [16]

$$\begin{aligned} \hat{\mathbf{s}} &= E[\mathbf{s}|\mathbf{y}] \\ &= \mathbf{C_s}(\mathbf{C_s} + \mathbf{C_v})^{-1}\mathbf{y}, \end{aligned} \tag{2}$$

where $\mathbf{C_s}$ and $\mathbf{C_v}$ are the covariance matrices of the signal and the noise, respectively. The covariance matrix is defined as $\mathbf{C_s} = E[\mathbf{s}\mathbf{s}^H]$, where $(\cdot)^H$ denotes Hermitian transposition and $E[\cdot]$ denotes the ensemble average operator.

3.2 Frequency domain LMMSE estimator and Wiener filter

In the frequency domain the goal is to estimate the complex DFT coefficients given a set of DFT coefficients of the noisy observation. Let $Y(m,k)$, $\theta(m,k)$, and $V(m,k)$ denote the m'th DFT coefficient of the k'th frame of the noisy observation, the signal, and the noise, respectively. Due to the linearity of the DFT operator, we have,

$$Y(m,k) = \theta(m,k) + V(m,k). \tag{3}$$

In vector form we have

$$\mathbf{Y} = \boldsymbol{\theta} + \mathbf{V}, \tag{4}$$

where again boldface letters represent vectors and the frame indices are omitted. As an example, the noisy spectrum vector of the k'th frame is arranged as

$$\mathbf{Y} = [Y(1,k), Y(2,k), \cdots, Y(N,k)]^T$$

where the number of frequency bins is equal to the number of samples per frame N.

We again use the linear model. $\mathbf{Y}$, $\boldsymbol{\theta}$, and $\mathbf{V}$ are assumed to be zero-mean complex Gaussian random variables and $\boldsymbol{\theta}$ and $\mathbf{V}$ are assumed to be uncorrelated to each other. The LMMSE estimate is the conditional mean

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= E[\boldsymbol{\theta}|\mathbf{Y}] \\ &= \mathbf{C}_\theta(\mathbf{C}_\theta + \mathbf{C}_\mathbf{V})^{-1}\mathbf{Y}, \end{aligned} \tag{5}$$

where $\mathbf{C}_\theta$ and $\mathbf{C}_\mathbf{V}$ are the covariance matrices of the DFT coefficients of the signal and the noise, respectively. By applying inverse DFT to each side, (5) can be easily shown to be identical to (2).

The relation between the two signal covariance matrices in time and frequency domain is

$$\mathbf{C}_\theta = \mathbf{F}\mathbf{C}_\mathbf{s}\mathbf{F}^{-1}, \tag{6}$$

where $\mathbf{F}$ is the Fourier matrix. If the frame was infinitely long and the signal was stationary, $\mathbf{C}_\mathbf{s}$ would be an infinitely large Toeplitz matrix. The infinite Fourier matrix is known to be the eigenvector matrix of any infinite Toeplitz matrix [8]. Thus, $\mathbf{C}_\theta$ becomes diagonal and the LMMSE estimator (5) reduces to the non-causal IIR Wiener filter with the transfer function

$$H_{WF}(\omega) = \frac{P_{ss}(\omega)}{P_{ss}(\omega) + P_{vv}(\omega)}, \tag{7}$$

where $P_{ss}(\omega)$ and $P_{vv}(\omega)$ denotes the power spectral density (PSD) of the signal and the noise, respectively. In the sequel we refer to (7) as the Wiener filter or WF.

4 High temporal resolution modeling for the signal covariance matrix estimation

For both time and frequency domain LMMSE estimators described in Section 3, the estimation of the signal covariance matrix $\mathbf{C_s}$ is crucial. In this work, we assume the noise to be stationary. For the signal, however, we propose the use of a high temporal resolution model to capture the non-stationarity caused by the excitation power variation. This can be explained by examining the voice production mechanism. In the well known source-filter model for voiced speech, the excitation source models the glottal pulse train, and the filter models the resonance property of the vocal tract. The vocal tract can be viewed as a slowly varying part of the system. Typically in a duration of 20 to 30 ms it changes very little. The vocal folds vibrate at a faster rate producing periodic glottal flow pulses. Typically there can be 2 to 8 glottal pulses in 20 ms. In speech coding, it is common practice to model this pulse train by a long-term correlation pattern parameterized by a long-term predictor [17] [18] [19]. However, this model fails to describe the linear relationship between the phases of the harmonics. That is, the long term predictor alone does not model the temporal localization of power in the excitation source. Instead, we apply a time envelope that captures the localization and concentration of pitch pulse energy in the time domain. This, in turn, introduces an element of non-stationarity to our signal model because the excitation sequence is now modeled as a random sequence with time varying variance, i.e., the glottal pulses are modeled with higher variance and the rest of the excitation sequence is modeled with lower variance. This modeling of non-stationarity within a short frame implies a temporal resolution much finer than that of the quasi-stationarity based algorithms. The latter has a temporal resolution equal to the frame length. Thus we term the former the high temporal resolution model. It is worth noting that some unvoiced phonemes, such as plosives, have very fast changing waveform envelopes, which also could be modeled as non-stationarity within the analysis frame. In this paper, however, we focus on the non-stationary modeling of voiced speech.

4.1 Modeling signal covariance matrix

The signal covariance matrix is usually estimated by averaging the outer product of the signal vector over time. As an example this is done in the signal subspace approach [10]. This method assumes ergodicity of the autocorrelation function within the averaging interval.

Here we propose the following method of estimating $\mathbf{C_s}$ with the ability to model a certain element of non-stationarity within a short frame. The following discussion is only appropriate for voiced speech. Let $\mathbf{r}$ denote the excitation source vector, and $\mathbf{H}$ denote the synthesis filtering matrix corresponding to the vocal tract filter such as

$$\mathbf{H} = \begin{bmatrix} h(0) & 0 & 0 & \cdots & 0 \\ h(1) & h(0) & 0 & & \vdots \\ h(2) & h(1) & h(0) & & \\ \vdots & \vdots & & \ddots & 0 \\ h(N-1) & h(N-2) & \cdots & & h(0) \end{bmatrix},$$

where $h(n)$ is the impulse response of the LPC synthesis filter. We then have

$$\mathbf{s} = \mathbf{H}\mathbf{r}, \tag{8}$$

and therefore

$$\mathbf{C_s} = E[\mathbf{s}\mathbf{s}^H] = \mathbf{H}\mathbf{C_r}\mathbf{H}^H, \tag{9}$$

where $\mathbf{C_r}$ is the covariance matrix of the model residual vector $\mathbf{r}$. In (9) we treat $\mathbf{H}$ as a deterministic quantity. This simplification is common practice also when the LPC filter model is used to parameterize the power spectral density in classic Wiener filtering [20] [5]. Section 4.2 addresses the estimation of $\mathbf{H}$. Note that (8) does not take into account the zero-input response of the filter in the previous frame. Either the zero-input response can be subtracted prior to the estimation of each frame, or a windowed overlap-add procedure can be applied to eliminate this effect.

We now model $\mathbf{r}$ as a sequence of independent zero mean random variables. The covariance matrix $\mathbf{C_r}$ is therefore diagonal with the variance of each element of $\mathbf{r}$ as its diagonal elements. For voiced speech, except for the pitch impulses, the rest of the residual is of very low amplitude and can be modeled as constant variance random variables. Therefore, the diagonal of $\mathbf{C_r}$ takes the shape of a constant floor with a few periodically located impulses. We term this the temporal envelope of the instantaneous residual power. This temporal envelope is an important part of the new MMSE estimator because it provides the information of uneven temporal power distribution. In the following two subsections, we will describe the estimation of the spectral envelope and the temporal envelope respectively.

4.2 Estimating the spectral envelope

In the context of LPC analysis, the synthesis filter has a spectrum that is the envelope of the signal spectrum. Thus, our goal in this subsection is to estimate the spectral envelope of the signal. We first use the Decision Directed method [3] to estimate the signal power spectrum and then use the autocorrelation method to find the spectral envelope.

The noisy signal power spectrum of the k'th frame $|\mathbf{Y}(k)|^2$ is obtained by applying the DFT to the k'th observation vector $\mathbf{y}(k)$ and squaring the amplitudes. The Decision Directed estimate of the signal power spectrum of the k'th frame, $|\hat{\hat{\boldsymbol{\theta}}}(k)|^2$, is a weighted

sum of two parts, the power spectrum of the estimated signal of the previous frame, $|\hat{\boldsymbol{\theta}}(k-1)|^2$, and the power-spectrum-subtraction estimate of the current frame's power spectrum:

$$|\hat{\hat{\boldsymbol{\theta}}}(k)|^2 = \alpha|\hat{\boldsymbol{\theta}}(k-1)|^2 + (1-\alpha)max(|\mathbf{Y}(k)|^2 - E[|\hat{\mathbf{V}}(k)|^2], 0), \qquad (10)$$

where α is a smoothing factor $\alpha \in [0,1]$, and $E[|\hat{\mathbf{V}}(k)|^2]$ is the estimated noise power spectral density. The purpose of such a recursive scheme is to improve the estimate of the power spectrum subtraction method by smoothing out the random fluctuation in the noise power spectrum, thus reduce the "musical noise" artifact [21]. Other iterative schemes with similar time or spectral constraints are applicable in this context. For a comprehensive study of constraint iterative filtering techniques, readers are referred to [5]. We now take the square-root of the estimated power spectrum and combine it with the noisy phase to reconstruct the so called intermediate estimate, which has the noise-reduced amplitude spectrum but noisy phase. An autocorrelation method LPC analysis is then applied to this intermediate estimate to obtain the synthesis filter coefficients.

4.3 Estimating the temporal envelope

We propose to use a modified MPLPC method to robustly estimate the temporal envelope of the residual power. MPLPC is first introduced by Atal and Remde [17] to optimally determine the impulse position and amplitude of the excitation in the context of analysis-by-synthesis linear predictive coding. The principle is to represent the LPC residual with a few impulses in which the locations and amplitudes (gains) of the impulses are chosen such that the difference between the target signal and the synthesized signal is minimized. In the noise reduction scenario, the target signal will be the noisy signal and the synthesis filter must be estimated from the noisy signal. Here, the synthesis filter is treated as known. For the residual of voiced speech, there is usually one dominating impulse in each pitch period. We first determine one impulse per pitch period, then model the rest of the residual as a noise floor with constant variance. In MPLPC the impulses are found sequentially [22]. The first impulse location and amplitude is found by minimizing the distance between the synthesized signal and the target signal. The effect of this impulse is subtracted from the target signal and the same procedure is applied to find the next impulse. Because this way of finding impulses does not take into account the interaction between the impulses, re-optimization of the impulse amplitudes is necessary every time a new impulse is found. The number of pitch impulses p in a frame is determined in the following way. p is first assigned an initial value equal to the largest number of pitch periods possible in a frame. Then p impulses are determined using the above mentioned method. Only the impulses with an amplitude larger than a threshold are selected as pitch impulses. In our experiment, the threshold

is set to 0.5 times the largest impulse amplitude in this frame. Having determined the impulses, a white noise sequence representing the noise floor of the excitation sequence is added into the gain optimization procedure together with all the impulses. We use a codebook of 1024 white Gaussian noise sequences in the optimization. The white noise sequence that yields the smallest synthesis error to the target signal is chosen to be the estimate of the noise floor. This procedure is in fact a multi-stage coder with p impulse stages and one Gaussian codebook stage, with a joint re-optimization of gains. Detailed treatment of this optimization problem can be found in [23]. After the optimization, we use a flat envelope equal to the square of the gain of the selected noise sequence to model the variance of the noise floor. Finally, the temporal envelope of the instantaneous residual power is composed of the noise floor variance and the squared impulses. When applied to noisy signals, the MPLPC procedure can be interpreted as a non-linear Least Square fitting to the noisy signal, with the impulse positions and amplitudes as the model parameters.

5 The algorithm

Having obtained the estimate of the temporal envelope of the instantaneous residual power and the estimate of the synthesis filter matrix, we are able to build the signal covariance matrix in (9). The covariance matrix is used in the time LMMSE estimator (2) or in the spectral LMMSE estimator (5) after being transformed by (6).

The noise covariance matrix can be estimated using speech absent frames. Here, we assume the noise to be stationary. For the time domain LMMSE estimator (2), if the noise is white, the covariance matrix $\mathbf{C_v}$ is diagonal with the noise variance as its diagonal elements. In the case of colored noise, the noise covariance matrix is no longer diagonal and it can be estimated using the time averaged outer product of the noise vector. For the spectral domain LMMSE estimator (5), $\mathbf{C_V}$ is a diagonal matrix with the power spectral density of the noise as its diagonal elements. This is due to the assumed stationarity of the noise[1]. In the special case where the noise is white, the diagonal elements all equal the variance of the noise.

We model the instantaneous power of the residual of unvoiced speech with a flat envelope. Here, voiced speech is referred to as phonemes that require excitation from the vocal folds vibration, and unvoiced speech consists of the rest of the phonemes. We use a simple voiced/unvoiced detector that utilize the fact that voiced speech usually has most of its power concentrated in the low frequency band, while unvoiced speech has a relatively flat spectrum within 0 to $4kHz$. Every frame is low pass filtered and

[1] In modeling the spectral covariance matrix of the noise we have ignored the inter-frequency correlations caused by the finite-length window effect. With typical window length, e.g. 15 to $30ms$, the inter-frequency correlations caused by the window effect is less significant than those caused by the non-stationarity of the signal. This can be easily seen by examining a plot of the spectral covariance matrix.

Algorithm 1 TFE-MMSE estimator

1: Take the k'th frame,
2: Estimate the noise PSD from the latest speech-absent frame.
3: Calculate the power spectrum of the noisy signal.
4: Do power spectrum subtraction estimation of the signal PSD, and refine the estimate using Decision-Directed smoothing (eq.(10)).
5: Reconstruct the signal by combining the amplitude spectrum estimated by 4 and the noisy phase.
6: Do LPC analysis to the reconstructed signal. Obtain the synthesis filter coefficients, and form the synthesis matrix $\mathbf{H}$.
7: **IF** the frame is voiced
   Estimate the envelope of the instantaneous residual power using the modified MPLPC method.
8: **IF** the frame is unvoiced
   Use a constant envelope for the instantaneous residual power.
9: **ENDIF**
10: Calculate the residual covariance matrix $\mathbf{C_r}$.
11: Form the signal covariance matrix $\mathbf{C_s} = \mathbf{H}\mathbf{C_r}\mathbf{H}^H$ (eq.(9)).
12: **IF** time domain LMMSE:
   $\hat{\mathbf{s}} = \mathbf{C_s}(\mathbf{C_s} + \mathbf{C_v})^{-1}\mathbf{y}$ (eq.(2)).
13: **IF** frequency domain LMMSE:
   transform $\mathbf{C_s}$ to frequency domain $\mathbf{C}_\theta = \mathbf{F}\mathbf{C_s}\mathbf{F}^{-1}$,
   filter the noisy spectrum $\hat{\boldsymbol{\theta}} = \mathbf{C}_\theta(\mathbf{C}_\theta + \mathbf{C_V})^{-1}\mathbf{Y}$ (eq.(5)),
   obtain the signal estimate by inverse DFT.
14: **ENDIF**
15: Calculate the power spectrum of the filtered signal, $|\hat{\boldsymbol{\theta}}(k-1)|^2$, for use in the PSD estimation of next frame.
16: $k = k + 1$ and go to 1.

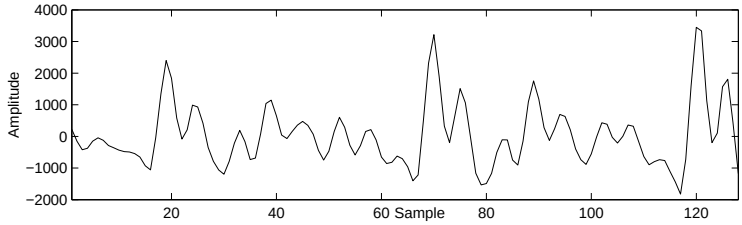
then the filtered signal power is compared with the original signal power. If the power loss is more than a threshold, the frame is marked as an unvoiced frame, and vice versa. Note however, that even for the unvoiced frames, the spectral covariance matrix is non-diagonal because the signal covariance matrix $\mathbf{C_s}$, built in this way, is not Toeplitz. Hereafter, we refer to the proposed approach as the Time-Frequency-Envelope MMSE estimator (TFE-MMSE), due to its utilization of envelopes in both time and frequency domain. The algorithm is summarized in Algorithm 1.

6 Reducing computational complexity

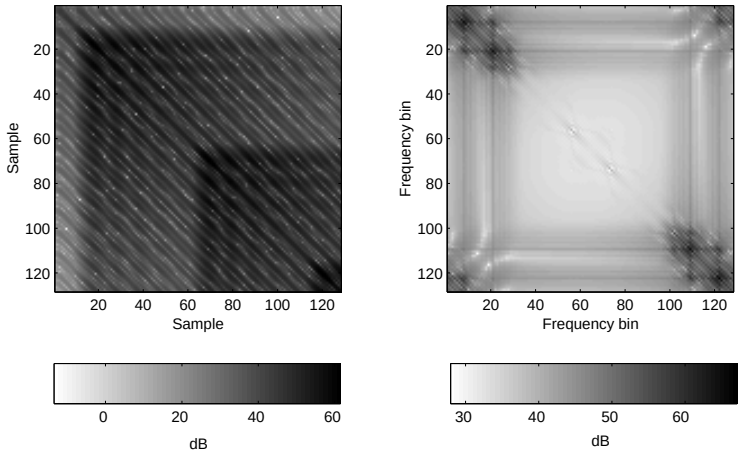
The TFE-MMSE estimators require inversion of a full covariance matrix $\mathbf{C_s}$ or $\mathbf{C}_\theta$. This high computational load prohibits the algorithm from real time application in hear-

ing aids. Noticing that both covariance matrices are symmetric and positive definite, Cholesky factorization can be applied to the covariance matrices, and the inversion can be done by inverting the Cholesky triangle. A careful implementation requires $N^3/3$ operations for the Cholesky factorization [24] and the algorithm complexity is $\mathcal{O}(N^3)$. Another computation intensive part of the algorithm is the modified MPLPC method. In this section we propose simplifications to these two parts.

Further reduction of complexity for the filtering requires understanding of the inter-frequency correlation. In the time domain the signal samples are clearly correlated with each other in a very long span. However, in the frequency domain, the correlation span is much smaller. This can be seen from the magnitude plots of the two covariance matrices (see Fig.1).



(a) Voiced speech waveform



(b) covariance matrix in time domain (left) and spectral domain (right)

Figure 1: The voiced speech waveform and its time domain and frequency domain (amplitude) covariance matrices estimated with the non-stationary model. Frame length is 128 samples.

For the spectral covariance matrix, the significant values concentrate around the diagonal. This fact indicates that a small number of diagonals capture most of the inter-

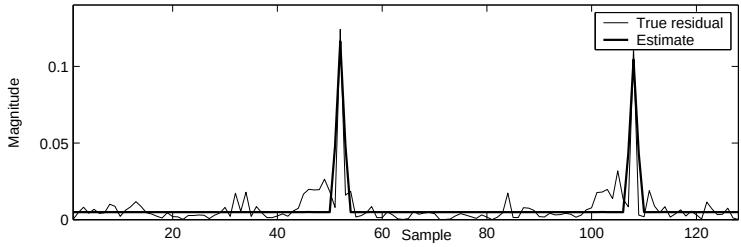
frequency correlation. The simplified procedure is as follows. Half of the spectrum vector $\boldsymbol{\theta}$ is divided into small segments of l frequency bins each. The sub-vector starting at the j'th frequency is denoted as $\boldsymbol{\theta}_{sub,j}$, where $j \in [1, l, 2l, \cdots, N/2]$ and $l \ll N$. The noisy signal spectrum and the noise spectrum can be segmented in the same way giving $\mathbf{Y}_{sub,j}$ and $\mathbf{V}_{sub,j}$. The LMMSE estimate of $\boldsymbol{\theta}_{sub,j}$ needs only a block of the covariance matrix, which means that the estimate of a frequency component benefits from its correlations with l neighboring frequency components instead of all components. This can be written as

$$\hat{\boldsymbol{\theta}}_{sub,j} = \mathbf{C}_{\theta_{sub,j}}(\mathbf{C}_{\theta_{sub,j}} + \mathbf{C}_{V_{sub,j}})^{-1}\mathbf{Y}_{sub,j}. \tag{11}$$

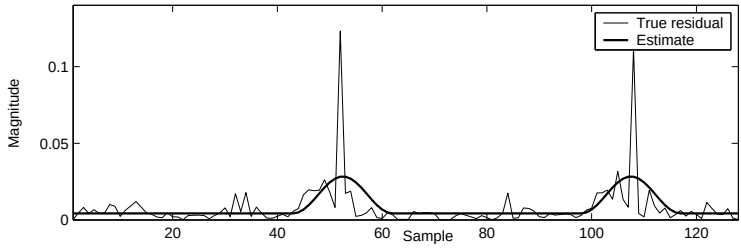
The first half of the signal spectrum can be estimated segment by segment. The second half of the spectrum is simply a flipped and conjugated version of the first half. The segment length is chosen to be $l = 8$, which in our experience does not degrade performance noticeably when compared with the use of the full matrix. Other segmentation schemes are applicable, such as overlapping segments. It is also possible to use a number of surrounding frequency components to estimate a single component at a time. We use the non-overlapping segmentation because it is computationally less expensive while maintaining good performance for small l. When the signal frame length is 128 samples and the block length is $l = 8$, using this simplified method requires only $\frac{8\times 8^3}{128^3} = \frac{1}{512}$ times of the original complexity for the filtering part of the algorithm with an extra expense of FFT operations to the covariance matrix. When l is set to values larger than 24, very little improvement in performance is observed. When l is set to values smaller than 8, the quality of enhanced speech degrades noticeably. By tuning the parameter l, an effective trade-off between the enhanced speech quality and the computational complexity is adjusted conveniently.

In the MPLPC part of the algorithm, the optimization of the impulse amplitude and the gain of the noise floor brings in heavy computational load. It can be simplified by fixing the impulse shape and the noise floor level. In the simplified version, the MPLPC method is only used for searching the locations of the p dominating impulses. Once the locations are found, a predetermined pulse shape is put at each location. An envelope of the noise floor is also predetermined. The pulse shape is chosen to be wider than an impulse in order to gain robustness against estimation error of the impulse locations. This is helpful as long as noise is present. The pulse shape used in our experiment is a raised cosine waveform with a period of 18 samples and the ratio between the pulse peak and the noise floor amplitude is experimentally determined to be 6.6. Finally, the estimated residual power must be normalized. Although the pulse shape and the relative level of the noise floor are fixed for all frames, experiments show that the TFE-MMSE estimator is not sensitive to this change. The performance of both the simplified procedure and the optimum procedure are evaluated in Section 7. Fig.2 shows the estimated envelopes

of residual in the two ways.



(a) Complete MPLPC method



(b) Simplified MPLPC method

Figure 2: Estimated magnitude envelopes of the residual by the MPLPC method and the simplified MPLPC method.

7 Results

Objective performance of the TFE-MMSE estimator is first evaluated and compared with the Wiener filter [2], the MMSE-LSA estimator [4], and the signal subspace method TDC estimator [10]. For the TFE-MMSE estimator, both the complete algorithm and the simplified algorithms are evaluated. For all estimators the sampling frequency is 8kHz, and the frame length is 128 samples with 50% overlap. In the Wiener filter we use the same Decision Directed method as in the MMSE-LSA and the TFE-MMSE estimator to estimate the PSD of the signal. An important parameter for the Decision Directed method is the smoothing factor α. The larger the α is, the more noise is removed and more distortion imposed to the signal, because of more smoothing made to the spectrum. In the MMSE-LSA estimator with the aforesaid parameter setting, we found experimentally $\alpha = 0.98$ to be the best trade-off between noise reduction and signal distortion. We use the same α for the WF and the TFE-MMSE estimator as for the MMSE-LSA estimator. For the TDC, the parameter μ ($\mu \geqq 1$) controls the degree of over suppression of the noise power [10]. The larger the μ is, the more attenuation

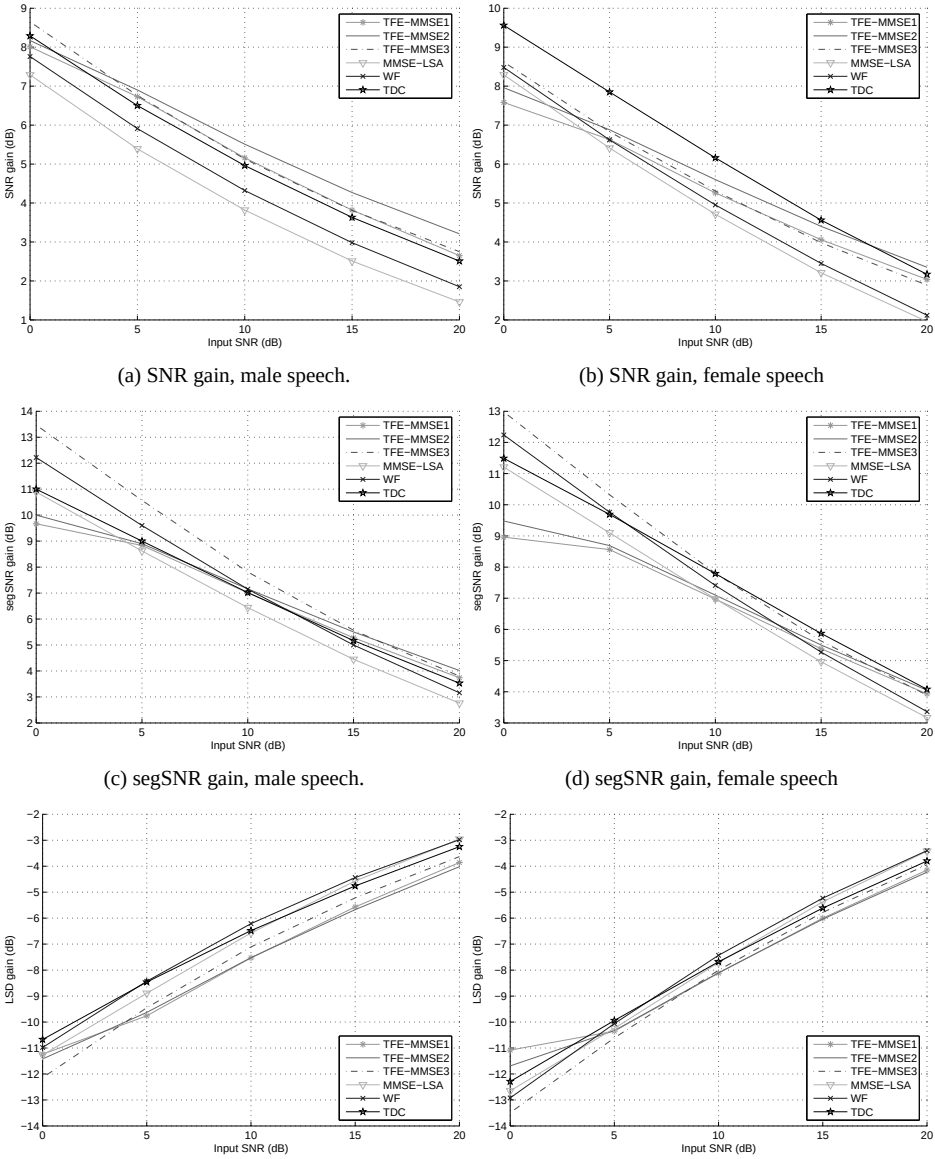
to the noise but larger distortion to the speech. We choose $\mu = 3$ in the experiments by balancing the noise reduction and signal distortion.

All estimators run with 32 sentences from different speakers (16 male and 16 female) from the TIMIT database [25] added with white Gaussian noise, pink noise, and car noise in SNR ranging from 0 dB to 20 dB. The white Gaussian noise is computer generated, and the pink noise is generated by filtering white noise with a filter having a 3 dB per octave spectral power descend. The car noise is recorded inside a car with a constant speed. Its spectrum is more low pass than the pink noise. The quality measures used include the SNR, the segmental SNR, and the Log-Spectral Distortion (LSD). The SNR is defined as the ratio of the total signal power to the total noise power in the sentence. The segmental SNR (segSNR) is defined as the average ratio of signal power to noise power per frame. To prevent the segSNR measure from being dominated by a few extreme low values, since the segSNR is measured in dB, it is common practice to apply a lower power threshold ϵ to the signals. Any frame that has an average power lower than ϵ is not used in the calculation. We set ϵ to 40dB lower than the average power of the utterance. The segSNR is commonly considered to be more correlated to perceived quality than the SNR measure. The LSD is defined as [26]:

$$LSD = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{M}\sum_{m=1}^{M}\left(20log_{10}\frac{|X(m,k)|+\epsilon}{|\hat{X}(m,k)|+\epsilon}\right)^2\right]^{\frac{1}{2}}, \tag{12}$$

where ϵ is to prevent extreme low values. We again set ϵ to 40 dB lower than the average power of the utterance. Results of the white Gaussian noise case are given in Fig. 3. TFE-MMSE1 is the complete algorithm, and TFE-MMSE2 is the one with simplified MPLPC and reduced covariance matrix ($l = 8$). It is observed that the TFE-MMSE2, although a result of simplification of TFE-MMSE1, has better performance than the TFE-MMSE1. This can be explained as follows: 1) Its wider pulse shape is more robust to the estimation error of impulse positions, and 2) the wider pulse shape can model to some extent the power concentration around the impulse peaks, which is overlooked by the spiky impulses. For this reason, in the following evaluations we investigate only the simplified algorithm.

Informal listening tests reveal that, although the speech enhanced by the TFE-MMSE algorithm has a significantly clearer sound (less muffled than the reference algorithms), the remaining background noise has musical tones. A solution to the musical noise problem is to set a higher value to the smoothing factor α. Using a larger α sacrifices the SNR and LSD slightly at high input SNR's, but improves the SNR and LSD at low input SNR's, and generally improves the segSNR significantly. The musical tones are also well suppressed. By setting $\alpha = 0.999$, the residual noise is greatly reduced, while the speech still sounds less muffled than for the reference methods. The reference methods can not use a smoothing factor as high as the TFE-MMSE: experiments show



(a) SNR gain, male speech.

(b) SNR gain, female speech

(c) segSNR gain, male speech.

(d) segSNR gain, female speech

(e) Log-Spectral Distortion gain, male speech.

(f) Log-Spectral Distortion gain, female speech.

Figure 3: SNR gain, segSNR gain, and Log-Spectral Distortion gain for the white Gaussian noise case.

that at $\alpha = 0.999$ the MMSE-LSA and the WF result in extremely muffled sounds. The TDC also suffers from a musical residual noise. To suppress its residual noise level to

as low as that of the TFE-MMSE with $\alpha = 0.999$, the TDC requires a μ lager than 8. This causes a sharp degradation of the SNR and LSD, and results in very muffled sounds. The TFE-MMSE2 estimator with a large smoothing factor ($\alpha = 0.999$) is hereafter termed TFE-MMSE3 and its objective measures are also shown in the figures. To verify the perceived quality of the TFE-MMSE3 subjectively, preference test between the TFE-MMSE3 and the WF, and between the TFE-MMSE3 and the MMSE-LSA are conducted. The WF and the MMSE-LSA use their best value of smoothing factor ($\alpha = 0.98$). The test is confined to white Gaussian noise and a limited range of SNR's. Three sentences by male speakers and three by female speakers at each SNR level are used in the test. Eight unexperienced listeners are required to vote for their preferred method based on the amount of noise reduction and speech distortion. The utterances are presented to the listeners by a high quality headphone. The clean utterance is first played as a reference, and the enhanced utterances are played once, or more if the listener finds this necessary. The results in Table 1 and 2 show that: 1) at 10 dB and 15 dB the listeners clearly prefer the TFE-MMSE over the two reference methods, while at 5 dB the preference on the TFE-MMSE is unclear; 2) the TFE-MMSE method has a more significant impact on the processing of male speech than on the processing of female speech. At 10 dB and above, the speech enhanced by TFE-MMSE3 has barely audible background noise, and the speech sounds less muffled than the reference methods. There is one artifact heard in rare occasions that we believe is caused by remaining musical tones. It is of very low power and occur some times at speech presence. The two reference methods have higher residual background noise and suffer from muffling and reverberance effects. When SNR is lower than 10 dB, a certain speech dependent noise occurs at speech presence in the TFE-MMSE3 processed speech. The lower the SNR is, the more audible this artifact is. Comparing the male and female speech processed by the TFE-MMSE3, the female speech sounds a bit rough.

The algorithms are also evaluated for pink noise and car noise cases. The objective results are shown in Fig. 4 and 5. In these results the TDC algorithm is not included because the algorithm is proposed based on the white Gaussian noise assumption. Informal listening test shows that the perceptual quality in the pink noise case for all the three algorithms are very similar to the white noise case, and that in the car noise case all tested methods have very similar perceptual quality due to the very low pass spectrum of the noise.

A comparison of spectrograms of a processed sentence (male "only lawyers love millionaires") is shown in Fig. 6.

Table 1: Preference test between WF and TFE-MMSE3 with additive white Gaussian noise.

		15 dB	10 dB	5 dB
Male speaker	WF	8%	7%	37%
	TFE	92%	83%	63%
Female speaker	WF	37%	33%	58%
	TFE	63%	67%	42%

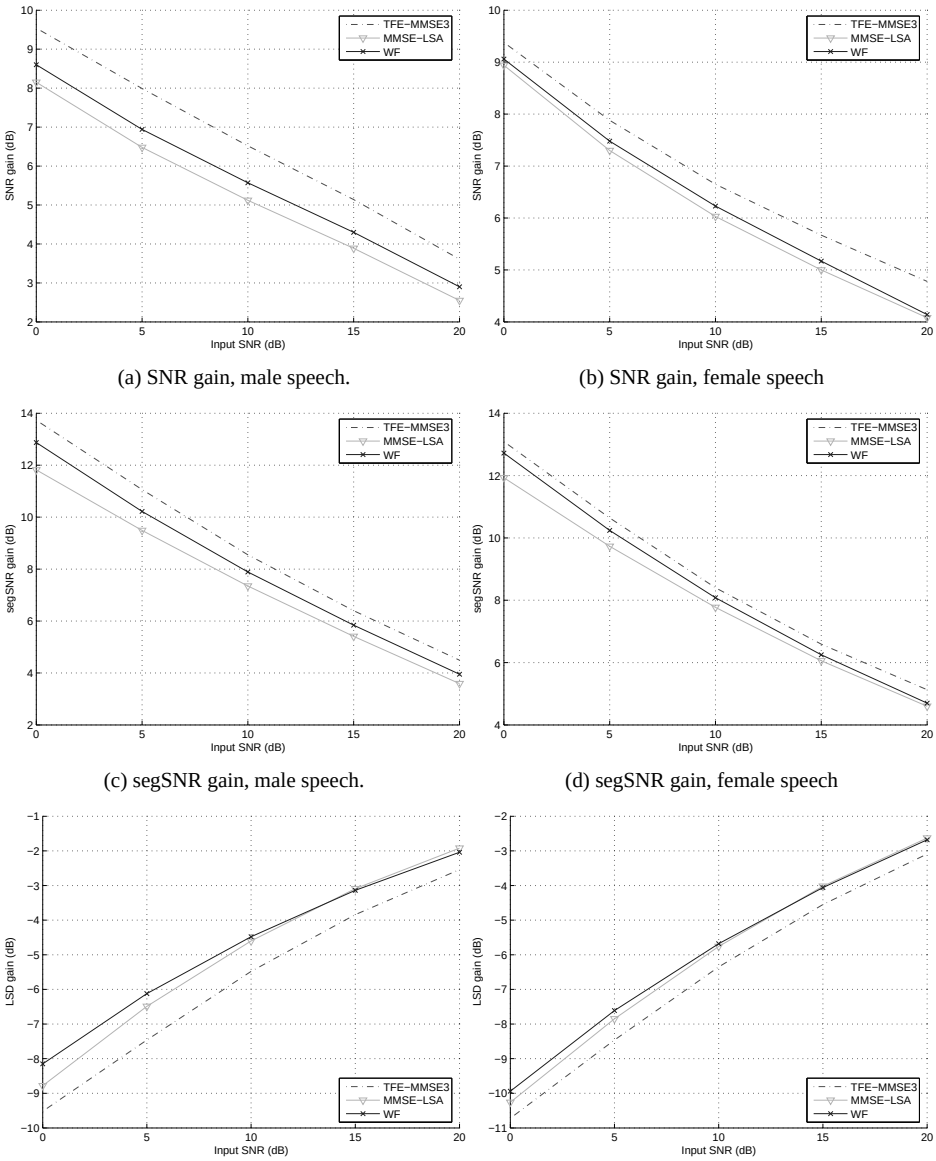
Table 2: Preference test between MMSE-LSA and TFE-MMSE3 with additive white Gaussian noise.

		15 dB	10 dB	5 dB
Male speaker	LSA	4%	25%	46%
	TFE	96%	75%	54%
Female speaker	LSA	25%	42%	50%
	TFE	75%	58%	40%

8 Discussion

The results show that for male speech, the TFE-MMSE3 estimator has the best performance in all the three objective measures (SNR, segSNR, and LSD). For female speech, the TFE-MMSE3 is the second in SNR, the best in LSD, and among the best in segSNR. The TFE-MMSE3 estimator allows a high degree of suppression to the noise while maintaining low distortion to the signal. The speech enhanced by the TFE-MMSE3 has a very clean background and a certain speech dependent residual noise. When the SNR is high (10 dB and above), this speech dependent noise is very well masked by the speech, and the resulting speech sounds clean and clear. As spectrograms in Fig. 6 indicates, the clearer sound is due to a better preserved signal spectrum, and a more suppressed background noise. At SNR lower than 5 dB, although the background still sounds clean, the speech dependent noise becomes audible, and perceived as a distortion to the speech.The listeners preference start shifting from the TFE-MMSE3 towards the MMSE-LSA that has a more uniform residual noise, although the noise level is high. The conclusion here is that at high SNR, it is preferable to remove background noise completely using the TFE-MMSE estimator without major distortion to the speech. This could be especially helpful at relieving listening fatigue for the hearing aid user. Whereas, at low SNR it is preferable to use a noise reduction strategy that produces uniform background noise, such as the MMSE-LSA algorithm.

The fact that female speech enhanced by the TFE-MMSE estimator sounds a little rougher than the male speech is consistent with the observation in [15], where male voiced speech and female voiced speech are found to have different masking properties in the auditory system. For male speech, the auditory system is sensitive to high frequency noise in the valleys between the pitch pulse peaks in the time domain. For the female speech, the auditory system is sensitive to low frequency noise in the valleys between the harmonics in the spectral domain. While the time domain valley for the



(a) SNR gain, male speech.

(b) SNR gain, female speech

(c) segSNR gain, male speech.

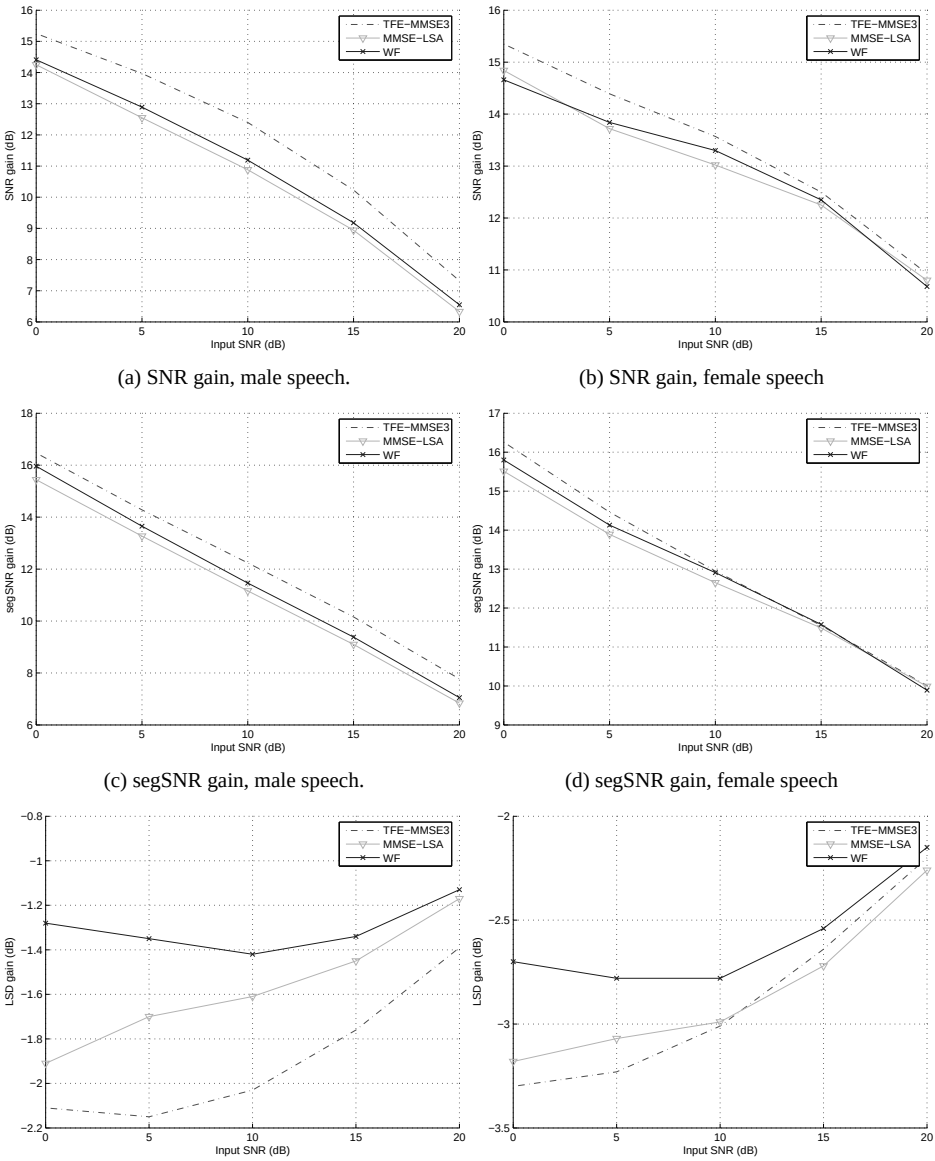
(d) segSNR gain, female speech

(e) Log-Spectral Distortion gain, male speech.

(f) Log-Spectral Distortion gain, female speech.

Figure 4: SNR gain, segSNR gain, and Log-Spectral Distortion gain for the pink noise case.

male speech is cleaned by the TFE-MMSE estimator, the spectral valleys for the female speech are not attenuated enough; a comb filter could help to remove the roughness in



(a) SNR gain, male speech.

(b) SNR gain, female speech

(c) segSNR gain, male speech.

(d) segSNR gain, female speech

(e) Log-Spectral Distortion gain, male speech.

(f) Log-Spectral Distortion gain, female speech.

Figure 5: SNR gain, segSNR gain, and Log-Spectral Distortion gain for the car noise case.

the female voiced speech.

In the TFE-MMSE estimator, we apply a high temporal resolution non-stationary

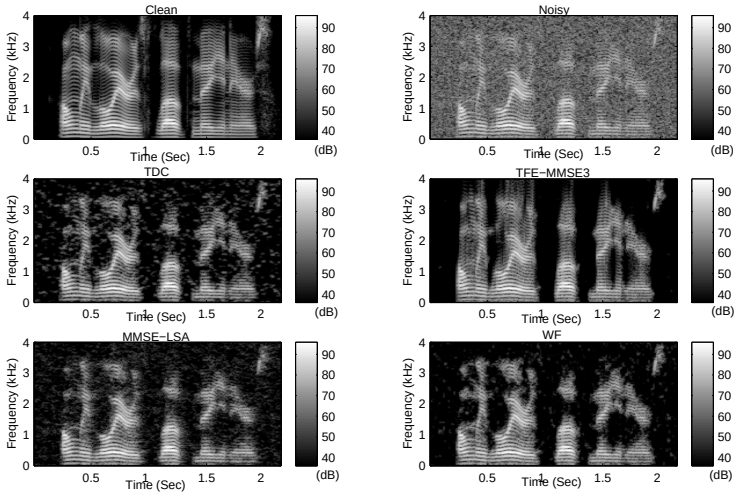


Figure 6: Spectrograms of enhanced speech. Input SNR is 10 dB.

model to explain the pitch impulses in the LPC residual of voiced speech. This enables the capture of abrupt changes in sample amplitude that are not captured by an AR linear stochastic model. In fact, the estimate of the residual power envelope contains information about the uneven distribution of signal power in time axis. In Fig.7 the original signal waveform, the WF enhanced signal waveform and the TFE-MMSE enhanced signal waveform of a voiced segment are plotted. It can be observed in this figure that by a better model of temporal power distribution the TFE-MMSE estimator represents the sudden rises of amplitude better than the Wiener filter.

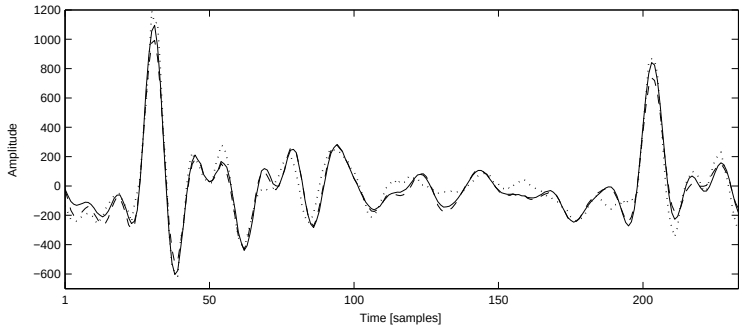


Figure 7: Comparison of waveforms of enhanced signals and the original signal. Dotted line: original, solid line: TFE-MMSE, dashed line: WF.

Noise in the phase spectrum is reduced by the TFE-MMSE estimator. Although

human ears are less sensitive to phase than to power, it is found in recent work [27] [28] [29] that phase noise is audible when the source SNR is very low. In [27] a threshold of phase perception is found. This phase-noise tolerance threshold corresponds to an SNR threshold of about 6 dB, which means for spectral components with local SNR smaller than 6 dB, it is necessary to reduce phase noise. The TFE-MMSE estimator has the ability of enhancing phase spectra because of its ability to estimate the temporal localization of residual power. It is the linearity in the phase of harmonics in the residual that makes the power be concentrated at periodic time instances, thus producing pitch pulses. Estimating the residual power temporal envelope enhances the linearity of the phase spectrum of the residual and therefore reduces phase noise in the signal.

References

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, No. 2, pp. 113–120, Apr. 1979.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[4] ——, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, pp. 443–445, Apr. 1985.

[5] J. H. L. Hansen and M. A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, 1991.

[6] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation With Gamma Distributed Speech Priors," *Proc.of ICASSP 2002*, vol. 1, pp. 253–256, May 2002.

[7] W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*. New York: McGraw-Hill, 1958.

[8] R. M. Gray, "Toeplitz and Circulant Matrices: A review," *Foundations and Trends in Communications and Information Theory*, vol. 2, Issue 3, pp. 155–239, 2006.

[9] C. Li and S. V. Andersen, "Inter-frequency Dependency in MMSE Speech Enhancement," *Proceedings of the 6th Nordic Signal Processing Symposium*, June 2004.

[10] Y. Ephraim and H. L. V. Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Tran. Speech and Audio Processing*, vol. 3, pp. 251–266, July 1995.

[11] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech Enhancement from Noise: A Regenerative Approach," *Speech Communication*, vol. 10, pp. 45–57, Feb. 1991.

[12] D. Tsoukalas, J. Mourjoupoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. on Speech and Audio Processing*, vol. 5(6), pp. 497–514, Nov. 1997.

[13] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. on Speech and Audio Processing*, vol. 7,no.2, pp. 126–137, 1999.

[14] K. Arehart, J. Hansen, S. Gallant, and L. Kalstein, "Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing impaired listeners," *Speech Communications*, vol. 40, no.4, pp. 575–592, Sept. 2003.

[15] J. Skoglund and W. B. Kleijn, "On Time-Frequency Masking in Voiced Speech," *IEEE Trans. Speech and Audio Processing*, vol. 8, No.4, pp. 361–369, July 2000.

[16] S. M. Kay, *Fundamentals of Statistical Signal Processing - Estimation Theory*. Prentice Hall PTR, 1993.

[17] B. Atal and J. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," *Proc. of ICASSP 1982*, vol. 7, pp. 614–617, May 1982.

[18] B. Atal, "Predictive Coding of Speech at Low Bit Rate," *IEEE Trans. on Comm.*, pp. 600–614, Apr. 1982.

[19] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *Bell Syst. Techn. J.*, vol. 49, pp. 1973–1986, 1970.

[20] J. S. Lim and A. V. Oppenheim, "All-pole Modeling of Degraded Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASP-26, pp. 197–209, June 1978.

[21] O. Cappé, "Elimination of the Musical Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2, pp. 345–349, Apr. 1994.

[22] A. M. Kondoz, *Digital Speech, Coding for Low Bit Rate Communications Systems*. John Wiley & Sons, 1999.

[23] N. Moreau and P.Dymarski, "Selection of excitation vectors for the CELP coders," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 1, pp. 29–41, January 1994.

[24] G. H. Golub and C. F. V. Loan, *Matrix Computations*. The Johns Hopkins University Press, 1996.

[25] "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," CD-ROM, NTIS, 1990.

[26] J.-M. Valin, J. Rouat, and F. Michaud, "Microphone array post-filter for seperation of simultaneous non-stationary source," *ICASSP 2004*, pp. I–221, 2004.

[27] P. Vary, "Noise Suppression By Spectral Magnitude Estimation - Mechanism and Theoretical Limits," *Signal Processing 8*, pp. 387–400, May 1985.

[28] H. Pobloth and W. B. Kleijn, "On Phase Perception in Speech," *Proc.of ICASSP 1999*, vol. 1, pp. 29–32, Mar. 1999.

[29] J. Skoglund, W. B. Kleijn, and P. Hedelin, "Audibility of Pitch-Synchronously Modulated Noise," *Speech Coding For Telecommunications Proceeding, IEEE*, vol. 7-10, pp. 51–52, Sept. 1997.

Paper C

Integrating Kalman Filtering and Multi-pulse Coding for Speech Enhancement with a Non-stationary Model of the Speech Signal

Chunjian Li and Søren Vang Andersen

The paper has been published in
Proceedings of the Thirty-eighth Annual Asilomar Conference on Signals, Systems, and Computers.
November 7 - November 10, 2004, Pacific Grove, California, USA.

The layout has been revised.

Abstract

In this paper, speech enhancement via Kalman filtering is considered. A non-stationary signal model for the speech signal is first described. This model consists of a slowly varying AR model and an excitation source that exhibits a rapidly time-varying variance. The AR model and the excitation model fit nicely into the Kalman filtering framework, fully exploiting the capability of the Kalman filter to process non-stationary signals in an LMMSE optimum manner. The AR-model coefficients are estimated by a decision-directed type Power Spectral Subtraction method followed by an LPC analysis. For the robust estimation of the rapidly time-varying excitation model in the presence of noise, we propose the use of a Multi-Pulse Linear Predictive Coding (MPLPC) based method. The Kalman filtering algorithm based on the non-stationary signal model is able to partially avoid the commonly used quasi-stationarity assumption of the speech. Therefore the non-stationarity of the signal is fully exploited in suppressing the noise power that is more stationary. Our experiments show that the Kalman filter with rapidly time-varying variance modeling using the proposed MPLPC based method brings significant performance improvement both when compared to a baseline Kalman filtering method with quasi-stationarity assumption and when compared to the well-known MMSE Log-Spectral Amplitude estimator (MMSE-LSA).

1 introduction

Kalman filters have been applied to speech enhancement in the last two decades. An early proposal can be dated back to Paliwal and Basu in the late 80's [1]. The Kalman filter can be seen as a generalization of the Wiener filter. It therefore has important properties that are superior to those of the Wiener filter. One of the most fundamental differences between the Wiener filter and the Kalman filter is the ability of the latter to accommodate non-stationary signals. However, most Kalman filters previously proposed for speech enhancement have not fully exploited this aspect. On the contrary, it is common practice to simply segment the speech into short frames and assume the signal to be stationary within each frame [1–3]. This is also known as the quasi-stationarity assumption. Thus, the modeling of signal non-stationarity in these methods is not significantly different from common practice for Wiener filtering [4] and Spectral Subtraction [5] based speech enhancement methods.

The speech signal is known to be non-stationary due to the movement of the articulators consisting of the vocal tract and the vocal folds. The short time processing usually segments signals into frames with length of about 20 ms. This temporal resolution is good enough to resolve the movement of the vocal tract, but not enough to resolve the movement of the vocal folds. Reducing the frame length is in general undesirable because it undermines the capability of averaging that every spectral estimator relies on.

Therefore, for voiced speech, a model with high temporal resolution is desired to fully exploit the non-stationarity of the signal.

A Kalman filter with modeling of non-stationarity is proposed by Popescu and Zeljkovic [6]. This filter aims at modeling non-stationarity of the noise but still assumes the speech to be stationary within the analysis frame. Lee et al. proposed an EM-based noise reduction approach [7], in which the excitation source of an AR filter is modeled as an outcome from one of two Gaussian processes. These processes differ by having a low and a high variance, respectively. This is in contrast to the single variance used in other proposed Kalman filters. Goh et al. proposed another EM-based algorithm with a voiced-unvoiced speech model that is able to model the periodicity or long-term correlation in the excitation of the voiced speech [8]. This model is still a quasi-stationary model since the long-term correlation alone can not model the temporal power concentration in the excitation source.

In this paper, we present a Kalman filter based approach with an explicit effort to estimate the time varying variance of the excitation source. This is achieved by modeling the excitation as a combination of sparse impulses and a noise component with low variance. To robustly identify the locations of these pulses, we propose the use of a modified Multi-Pulse Linear Predictive Coding (MPLPC) method, which was originally proposed for lossy compression of speech by Atal and Remde [9]. The AR parameters are estimated in a recursive manners similar to the decision directed method in [10]. A forward-backward Kalman filtering using the estimated high temporal resolution excitation variance and the AR model is then applied to obtain a final estimate of the signal.

2 Non-stationary signal modeling

In [11] we show that voiced speech can be advantageously modeled as non-stationary even within a short analysis frame. Examining the speech production mechanism reveals that for voiced speech the vocal tract filter is slowly varying while the excitation source produced by the vocal folds exhibits rapid variation in power. An all-pole filter estimated by the Linear Predictive Coding (LPC) method excited by the LPC residual is a good mathematical model of speech production. With this model, the high temporal resolution estimation and robust spectral envelope estimation are divided into separate problems: the LPC residual exhibits rapid power variation, thus requires a high temporal resolution modeling; the all-pole filter represents the spectral envelope of the signal, thus demands large data length for a robust estimation. Therefore, our non-stationary signal model consists of an all-pole filter that is invariant within the span of a frame, and an excitation sequence modeled by N Gaussian random variables with zero means and varying variances, where N is the frame length. This is different from the quasi-stationary model, which models the excitation source as having a constant variance

within a frame. This signal model partially avoids the quasi-stationarity assumption, therefore is termed non-stationary signal model.

3 Kalman filtering

The non-stationary signal model is most suitable for Kalman filtering because of the Kalman filter's capability to handle non-stationarity. To fully utilize the data buffered in frames, as is the case in many applications, we choose to use a forward-backward Kalman filtering formulation.

We use the following state space model:

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{A}\mathbf{x}(n-1) + \mathbf{b}u(n) \\ y(n) &= \mathbf{h}\mathbf{x}(n) + v(n), \end{aligned} \tag{1}$$

where $\mathbf{x}$ is the state vector of the speech signal, $u(n)$ is the process noise, $y(n)$ is the observation, $v(n)$ is the observation noise, $\mathbf{A}$ is the state transition matrix, and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix}, \tag{2}$$

$$\mathbf{b}^T = \mathbf{h} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}. \tag{3}$$

The Kalman forward filtering solution is summarized as follows:

$$\hat{\mathbf{x}}(n|n-1) = \mathbf{A}\hat{\mathbf{x}}(n-1|n-1) \tag{4}$$

$$\mathbf{M}(n|n-1) = \mathbf{A}\mathbf{M}(n-1|n-1)\mathbf{A}^T + \sigma_u^2(n)\mathbf{b}\mathbf{b}^T \tag{5}$$

$$\mathbf{K}(n) = \frac{\mathbf{M}(n|n-1)\mathbf{h}^T(n)}{\sigma_v^2 + \mathbf{h}(n)\mathbf{M}(n|n-1)\mathbf{h}^T} \tag{6}$$

$$\begin{aligned} \hat{\mathbf{x}}(n|n) = {} & \hat{\mathbf{x}}(n|n-1) + \\ & \mathbf{K}(n)[y(n) - \mathbf{h}(n)\hat{\mathbf{x}}(n|n-1)] \end{aligned} \tag{7}$$

$$\mathbf{M}(n|n) = [\mathbf{I} - \mathbf{K}(n)\mathbf{h}(n)]\mathbf{M}(n|n-1). \tag{8}$$

The backward filtering solution can be shown to be [12]:

$$\hat{\mathbf{x}}(n-1|N) = \hat{\mathbf{x}}(n-1|n-1) + \mathbf{F}(n-1)[\hat{\mathbf{x}}(n|N) - \hat{\mathbf{x}}(n|n-1)] \quad (9)$$

$$\mathbf{F}(n-1) = \mathbf{M}(n-1|n-1)\mathbf{A}^T\mathbf{M}^{-1}(n|n-1). \quad (10)$$

In the equations listed above, $\hat{\mathbf{x}}(n|n-1)$ denotes the forward prediction of $\mathbf{x}(n)$ using previous data up to time $n-1$, and $\hat{\mathbf{x}}(n|n)$ denotes the forward filtering estimate using data up to time n. Likewise, $\mathbf{M}(n|n-1)$ and $\mathbf{M}(n|n)$ are the forward prediction and filtering estimate MSE matrix, respectively. The vector $\hat{\mathbf{x}}(n-1|N)$ denotes the backward prediction of $\mathbf{x}(n-1)$ using future data from time n to the end of the frame. The matrix $\mathbf{F}(n-1)$ denotes the backward prediction MSE matrix. The filtering first goes forward obtaining the forward estimate and forward MSE matrix, then goes backward and combine the forward-backward estimate by eq.(9). The unknown parameters need to be estimated before the filtering, which includes $\mathbf{A}$, $\sigma_u^2(n)$, and σ_v^2. The observation noise are assumed to be white Gaussian in this work. Its variance σ_v^2 is time invariant and can be estimated using the speech absent frames. The variance of the processing noise, on the other hand, is time varying. The estimation of $\mathbf{A}$ and $\sigma_u^2(n)$ will be presented in the following section.

4 Parameter estimation

4.1 AR parameter estimation

The estimate of AR coefficients is needed in building the state transition matrix of the Kalman filter. Since the AR model represents the spectral envelope of the signal, it is convenient to estimate the signal spectrum first and then estimate its envelope. To estimate the signal spectrum robustly and efficiently, we use the Power Spectral Subtraction method in a time-recursive manner similar to the decision directed method used in [13]. Denote the DFT spectrum of the speech in the kth frame by a vector $\boldsymbol{\theta}(k)$. The current estimate of the signal power spectrum of the k'th frame, $|\hat{\hat{\boldsymbol{\theta}}}(k)|^2$, is a weighted sum of two parts, the power spectrum of the estimated signal of the previous frame and the power-spectral-subtraction estimate of the current frame's power spectrum:

$$|\hat{\hat{\boldsymbol{\theta}}}(k)|^2 = \alpha|\hat{\boldsymbol{\theta}}(k-1)|^2 + (1-\alpha)max(|\mathbf{Y}(k)|^2 - E[|\hat{\mathbf{V}}(k)|^2], 0), \quad (11)$$

where α is a smoothing factor, $|\mathbf{Y}(k)|^2$ is the noisy power spectrum of k'th frame, $|\hat{\boldsymbol{\theta}}(k-1)|^2$ is the power spectrum of the estimated signal of the $(k-1)$'th frame and

$E[|\hat{\mathbf{V}}(k)|^2]$ is the estimated noise power spectral density. The smoothing factor α controls the degree of smoothing over time. Such a smoothing scheme has been shown to be effective in reducing musical noise artifact. We take the square-root of the estimated signal power spectrum and combine it with the noisy phase to obtain an intermediate estimate of the signal. An auto-correlation type LPC analysis is then applied to the intermediate estimate to obtain the estimate of the AR model coefficients.

4.2 Estimating the excitation variance with high temporal resolution

The conventional quasi-stationarity based algorithms estimate the excitation source variance by explicitly or implicitly averaging the power of the estimate of the excitation source over the whole frame. In our non-stationary model, in order to resolve the rapid power variation of the excitation of the voiced speech, the variance must be estimated within smaller intervals. Acknowledging the impulse train structure of the LPC residual (see Figure 1), a time varying variance can be found by first estimating the residual instantaneous power and then doing smoothing to the instantaneous power with less smoothing around the impulses and more smoothing between the impulses. The smoothed instantaneous power is our estimate of the variance. In this way, the onset of the power rise at the impulse is preserved, and the variance estimate between the impulses are robust to outliers because of higher degree of smoothing. Since impulses are of high amplitudes and easier to estimate than the floor between the impulses, when noise is present, we propose the following simplified procedure, which does not require estimating all samples of the excitation source. The positions of the impulses are first estimated, then a pre-determined pulse shape is put on every impulse position. A constant noise floor with an amplitude that is proportional to the pulse peak is put on, together with the pulses, to form an envelope of the instantaneous power of the excitation. The envelope is finally scaled to ensure that its total energy equals the estimated energy of the excitation. The pulse shape and the amplitude ratio are determined by experiments. We choose a raised cosine waveform with a period of 18 samples as the pulse shape, and the amplitude ratio is set to 6.6. To robustly estimate the impulse positions, we propose to use the Multi-Pulse Linear Predictive Coding (MPLPC) method. The basic MPLPC method is originally proposed by Atal and Remde [9] for determining the impulse position and amplitude of the excitation in linear predictive coding (LPC) applications. The MPLPC procedure finds the optimum position and amplitude of the excitation impulses that minimize the distance between the target signal waveform and the synthesized signal waveform. In our noise reduction application, the target signal is the noisy speech signal. The impulses are estimated in a sequential way: every time an impulse has been determined, its contribution to the waveform is subtracted and a search for the next impulse is started. The search continues until the amplitude of the newest impulse gets

below a certain threshold. We choose the threshold to be 0.5 times the highest impulse amplitude. Any new impulse smaller than this threshold is not regarded as a pitch impulse. The following is a brief description of the MPLPC optimization procedure. For details the reader is referred to [14].

The squared error between the synthesized signal using the first impulse and the noisy signal can be written as

$$e = \sum_{n=1}^{N} [y(n) - gh(n-m)]^2, \tag{12}$$

where N is the frame length, g and m are the amplitude and location of the impulse respectively, and $h(n)$ is the impulse response of the synthesis filter. By differentiating (12) with respect to g and setting the derivative to zero, the optimum amplitude is found to be

$$g = \frac{\sum_{n=1}^{N} y(n)h(n-m)}{\sum_{n=1}^{N} h^2(n-m)} \tag{13}$$

and the optimum value for m can be shown to be

$$m^* = \underset{m}{arg\,max} \frac{(\sum_{n=1}^{N} y(n)h(n-m))^2}{\sum_{n=1}^{N} h^2(n-m)}, \tag{14}$$

where m^* denotes the optimum position of the impulse. After the estimation of all the pitch impulses sequentially, only the position information is used in constituting the envelope, as described previously. An example of the estimated envelope is shown in Figure 1.

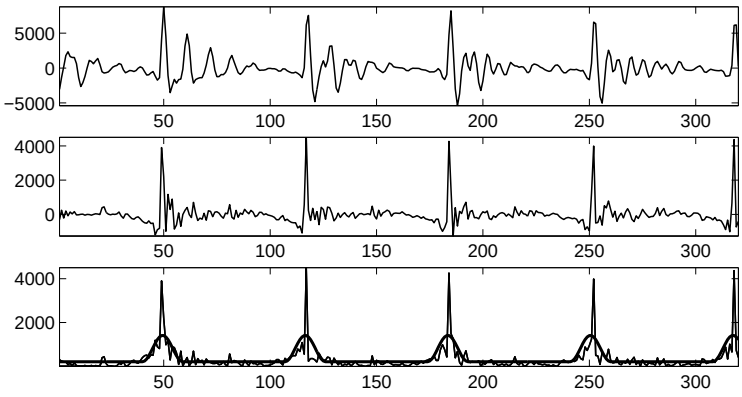


Figure 1: Top: a segment of voiced speech waveform; middle: the LPC residual of the speech waveform; bottom: the instantaneous magnitude of the residual (thin line) and the estimated amplitude envelope (thick line).

5 experimental results

To evaluate the performance of the proposed Non-stationary Kalman filter (NSK), we compare it with two reference methods, the conventional Kalman filter (CK) based on quasi-stationary assumption, and the MMSE-LSA estimator [10]. In CK, the all-pole model is estimated by the decision directed power subtraction method as same as the one used in the MMSE-LSA, followed by an LPC analysis. The smoothing factor α in all three algorithms is set to 0.98. All algorithms run with 32 sentences from the TIMIT database corrupted by white Gaussian noise at different SNR. The sampling frequency is 8 kHz and the frame length is 128 samples with 50% overlap. The comparison is on objective measures including SNR gain and Log-Spectral Distortion (LSD). The SNR is defined as the ratio between the total signal power and the noise power. The LSD is defined as the distance between log-scaled DFT spectra for the clean and the processed speech summed over all frequencies and divided by the number of frequency bins. Comparison of spectrograms, and informal listening test are also performed. Figure 2 and 3 show the results for SNR gain and LSD, respectively.

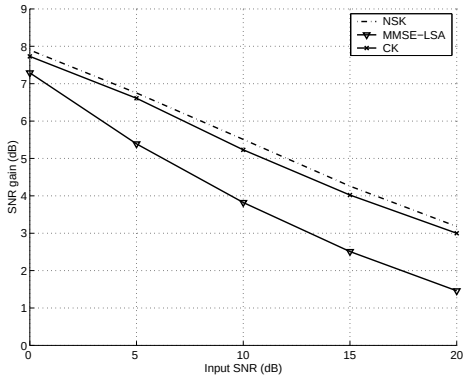


Figure 2: Comparison of SNR gain.

It is observed that the proposed NSK has constantly the highest SNR gain among the three algorithms, and has the lowest spectral distortion except for 0 dB input SNR. In Figure 4, the spectrograms of the processed speech by the three algorithms are compared. Here we clearly see that the NSK preserves the harmonic structure of the voiced speech better than all the other algorithms. Finally, informal listening test shows that the NSK results in a less muffled sound than the other two algorithms, as is evident from the spectrogram plots.

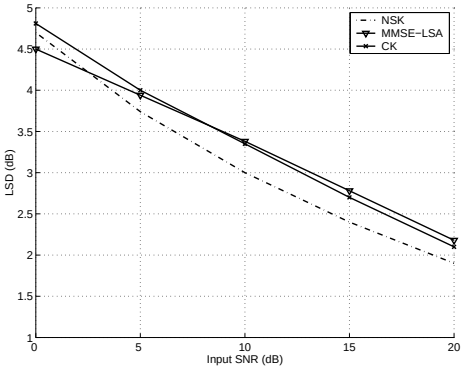


Figure 3: Comparison of LSD.

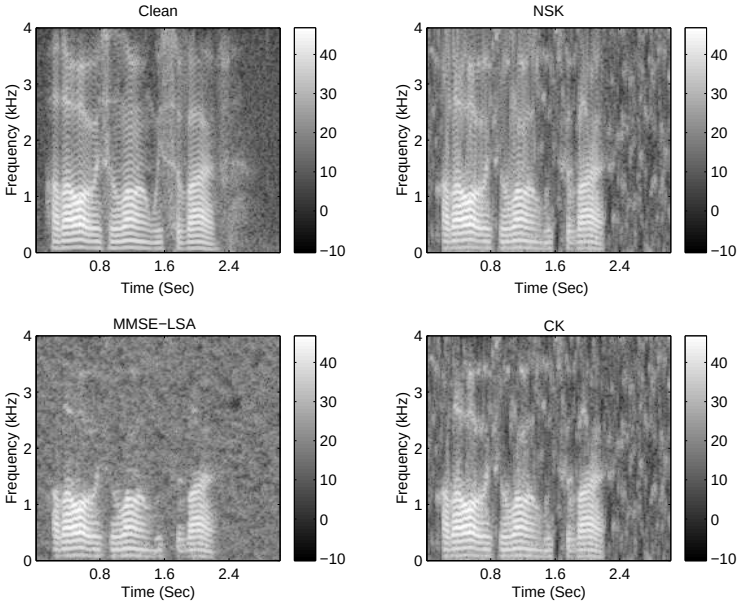


Figure 4: Comparison of spectrograms.

6 Conclusion

In this paper, we proposed a non-stationary signal model that is able to model the rapid power variation in the excitation source of the voiced speech signals. This model estimates the variance of the excitation source with a high temporal resolution by fitting an envelope to the instantaneous power of the LPC residual. The envelope is designed

to emphasize the temporal power concentration at the impulses while reducing noise power between the impulses. Locating the impulses is done by an MPLPC optimization procedure. The Kalman filter with this non-stationary signal model shows better SNR gain and suffers from lower spectral distortion than the quasi-stationarity based Kalman filter and MMSE-LSA estimator.

References

[1] K. K. Paliwal and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering," *Proc.of ICASSP 1987*, vol. 12, pp. 177–180, Apr. 1987.

[2] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement," *IEEE Trans. on Signal Processing*, vol. 39, pp. 1732–1742, 1991.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. on Speech and Audio*, vol. 6, pp. 373–385, July 1998.

[4] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[5] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, No. 2, pp. 113–120, Apr. 1979.

[6] D. C. Popescu and I. Zeljkovic, "Kalman filtering of colored noise for speech enhancement," *Proc. ICASSP*, vol. 2, pp. 997–1000, 1998.

[7] B. G. Lee, K. Y. Lee, and S. Ann, "An EM-based approach for parameter enhancement with an application to speech signals," *Signal Processing*, vol. 46, pp. 1–14, 1995.

[8] Z. Goh, K. Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. on Speech and Audio Processing*, vol. 7, No.5, pp. 510–524, 1999.

[9] B. Atal and J. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," *Proc. of ICASSP 1982*, vol. 7, pp. 614–617, May 1982.

[10] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, pp. 443–445, Apr. 1985.

[11] C. Li and S. V. Andersen, "Inter-frequency Dependency in MMSE Speech Enhancement," *Proceedings of the 6th Nordic Signal Processing Symposium*, June 2004.

[12] H. Rauch, "Solutions to the linear smoothing problem," *IEEE Trans. on Automatic Control*, vol. AC-8, 1963.

[13] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[14] A. M. Kondoz, *Digital Speech, Coding for Low Bit Rate Communications Systems*. John Wiley & Sons, 1999.

Paper D

A New Iterative Speech Enhancement Scheme Based on Kalman Filtering

Chunjian Li and Søren Vang Andersen

The paper has been published in
Proceedings of the 13th European Signal Processing Conference.
September 9-11, 2005, Antalya, Turkey.

The layout has been revised.

Abstract

A new iterative speech enhancement scheme that can be seen as an approximation to the Expectation-Maximization (EM) algorithm is proposed. The algorithm employs a Kalman filter that models the excitation source as a spectrally white process with a rapidly time-varying variance, which calls for a high temporal resolution estimation of this variance. A Local Variance Estimator based on a Prediction Error Kalman Filter is designed for this high temporal resolution variance estimation. To achieve fast convergence and avoid local maxima of the likelihood function, a Weighted Power Spectral Subtraction filter is introduced as an initialization procedure. Iterations are then made sequential inter-frame, exploiting the fact that the AR model changes slowly between neighboring frames. The proposed algorithm is computationally more efficient than a baseline EM algorithm due to its fast convergence. Performance comparison shows significant improvement over the baseline EM algorithm in terms of three objective measures. Listening test indicates an improvement in subjective quality due to a significant reduction of musical noise compared to the baseline EM algorithm.

1 Introduction

Single channel noise reduction of speech signals using iterative estimation methods has been an active research area for the last two decades. Most of the known iterative speech enhancement schemes are based on, or can be interpreted as, the Expectation-Maximization (EM) algorithm or a certain approximation to it. Proposals of the EM algorithms for speech enhancement can be found in [1] [2] [3] [4] [5]. Some other iterative speech enhancement techniques can be seen as approximations to the EM algorithm, see e.g. [6] [7] [8] [9]. A paradigm of these EM based approaches is to iterate between an expectation step comprising Wiener or Kalman filtering given the current estimate of signal model parameters, and a maximization step comprising the estimation of the parameters given the filtered signal. By doing so, the conditional likelihood of the estimated parameters and the signal increases monotonically until a certain convergence criterion is reached.

Evolution of these EM approaches is seen in the underlying signal models. In early proposals [6] [1] [7], the non-causal IIR Wiener filter (WF) is used, where the signal is modeled as a short-time stationary Gaussian process. This is a rather simplified model, where the speech is assumed to be stationary and the voiced and unvoiced speech share the same Gaussian model even though voiced speech is known to be far from Gaussian. The time domain formulation in [2] uses the Kalman smoother in place of the WF, which allows the signal to be modeled as non-stationary but still uses one model for both voiced and unvoiced speech. In [3], the speech excitation source is modeled as a mixture of two Gaussian processes with differing variances. For voiced speech, the process with

higher variance models the impulses and the one with lower variance models the rest of the excitation sequence. The detection of the impulse is done by a likelihood test at every time instant. In [4], an explicit model of speech production is used, where the excitation of voiced speech is modeled as an impulse train superimposed in white noise. The impulse parameters (pitch period, amplitude, and phase) and the noise floor variance are estimated iteratively by an inner loop in every iteration. In [9], the long term correlation in voiced speech is explicitly modeled. To accomplish this, the instantaneous pitch period and the degree of voicing need to be estimated in every frame. In general, using finer models has the potential to improve the enhanced speech quality, but also raises the concern of complexity and robustness, since the decision on voicing and other pitch related parameters are difficult to extract from noisy observations.

Another line of development in speech enhancement employing fine models of the voiced speech production mechanism puts effort into modeling the rapidly varying variance of the excitation source of voiced speech signals under a Linear Minimum Mean Squared-Error Estimator (LMMSE) framework [10] [11] [12]. It is shown that the prominent temporal localization of power in the excitation source of voiced speech is a major source of correlation between spectral components of the signal. An LMMSE estimator with a signal model that models this non-stationarity can achieve both higher SNR gain and lower spectral distortion. It is well known that the Kalman filter provides a more convenient framework for modeling signal non-stationarity than the WF: the WF assumes the signal to be wide-sense stationary; while the Kalman filter allows for a dynamic mean, which is modeled by the state transition model, and a dynamic system noise variance, which is assumed to be known *a priori*. Whereas, in most of the proposed Kalman filtering based speech enhancement approaches, the system noise variance is modeled as constant within a short frame, thus an important part of the non-stationarity is not modeled. In [12], the temporal localization of power in the excitation source is estimated by a modified Multi-pulse LPC method, and the Kalman filter using this dynamic system noise variance gives promising results.

In this paper, we propose a new iterative approach employing Kalman filtering with a signal model comprising a rapidly time-varying excitation variance. The proposed algorithm consists of three steps in every iteration, i.e., the estimation of the auto-regressive (AR) parameters, the excitation source variance estimation with high temporal resolution, and the Kalman filtering. The high temporal resolution estimation of the excitation variance is performed by a combination of a prediction-error Kalman filter and a spline smoothing method. By employing an initialization procedure called Weighted Spectral Power Subtraction, the convergence is achieved in one iteration per frame. The iterative scheme thus becomes frame-wise sequential, because the estimation in the current frame is based on the filtered signal of the previous frame. In constrast with the aforementioned EM approaches with fine speech production models, this approach has the advantages of simplicity and robustness since it requires no ex-

plicit estimation of pitch related parameters neither voiced/unvoiced decisions. The low computational complexity is also attributed to its fast convergence.

2 The Kalman filter based iterative scheme

It is convenient to introduce the overall scheme before going into detailed discussion. Figure 1 shows the function blocks of the proposed algorithm. The noisy signal is segmented into non-overlapping short analysis frames. We denote the nth sample of the speech signal, the additive noise, and the noisy observation of the kth frame as $s(n,k)$, $v(n,k)$ and $y(n,k)$, respectively. At the first iteration of the kth frame, the noisy signal is first filtered by a Weighted Power Spectral Subtraction (WPSS) filter as an initialization step. The WPSS does a Power Spectral Subtraction (PSS) estimation of the signal spectrum, and combines it with the estimated power spectrum of the previous frame. The filtered signal $\hat{s}_{pss}(n,k)$ is then synthesized using the combined spectrum and the noisy phase, and is fed into an LPC analysis (by closing the switch to the WPSS output) to estimate the AR coefficients. A Prediction Error Kalman filter (PEKF) takes the $\hat{s}_{pss}(n,k)$ as input and estimates the system noise $\hat{u}(n,k)$. The time dependent variance of the excitation, $\sigma_u^2(n,k)$, is estimated by a Local Variance Estimator (LVE) that locally smoothes the instantaneous power of the $\hat{u}(n,k)$. A second Kalman filter then filters the noisy signal to get the final signal estimate, using the estimated SR coefficients and system noise variance. The signal estimate $\hat{s}(n,k)$ is used by the LPC block in the next iteration (by closing the switch to the feed back link) to improve the estimation of the AR coefficients.

The iterations can be made sequential on a frame-to-frame basis by fixing the number of iterations to one, and closing the switch to the WPSS permanently. This is a frame-wise-sequential approximation to the original iterative algorithm, with the purpose of reducing computational complexity, exploiting the fact that the spectral envelope of the speech signal changes slowly between neighboring frames. As is shown in the experiment section, with an appropriate parameter setting of the WPSS procedure, the iterative algorithm can achieve convergence in the first iteration with an even higher SNR gain. For comparison, the block diagram of the iterative-batch EM approach (IEM) [2] [5] that is used as a baseline algorithm in our work is shown in Figure 2 (A). Note that for the IEM, the system noise variance is only dependent on the frame index k, while for the proposed algorithm, it is dependent on both k and n. The two new functional blocks in the proposed algorithm are the WPSS and the High Temporal Resolution Modeling (HTRM) block. The function of the WPSS is to improve the initialization of the iterative scheme to achieve fast convergence. Section 3 addresses the initialization issue in details. The HTRM block estimates the system noise variance in a high temporal resolution, in contrast to the IEM where the system noise variance is

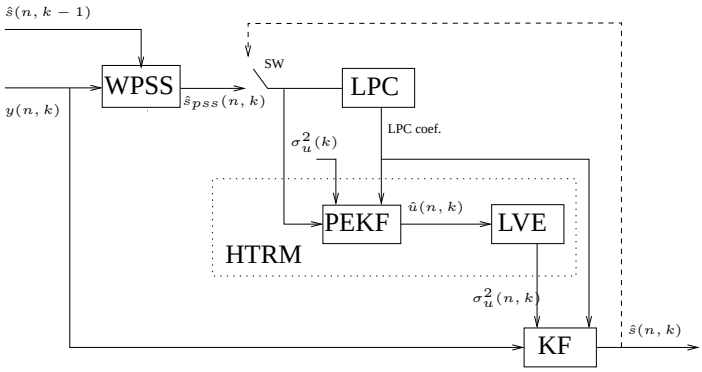


Figure 1: Block diagram of the proposed algorithm

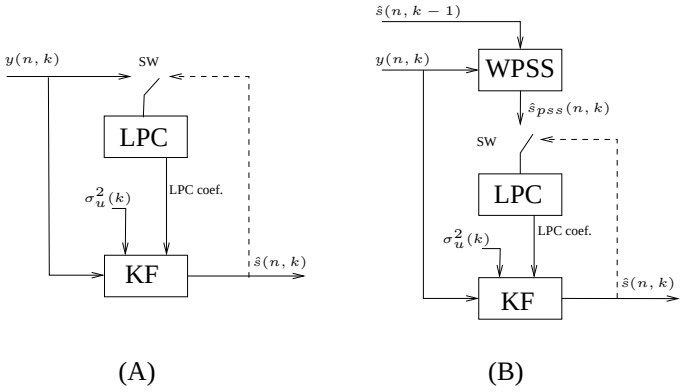


Figure 2: Block diagrams of the IEM algorithm (A), and the IEM with WPSS initialization (B) .

constant within a frame. The formulation of the Kalman filtering with high temporal resolution modeling is treated in section 4.

3 Initialization and sequential approximation

The Weighted Power Spectral Subtraction procedure combines the signal power spectrum estimated in the previous frame and the one estimated by the Power Spectral Subtraction method in the current frame, so that the iteration of the current frame is started with the result of the previous iteration as well as the new information in the current frame. The weight of the previous frame is set much larger than the weight of the current frame because the signal spectrum envelope varies slowly between neighboring

frames. The WPSS combines the spectrum estimates as follows:

$$|\hat{\hat{\boldsymbol{\theta}}}(k)|^2 = \alpha|\hat{\boldsymbol{\theta}}(k-1)|^2 + (1-\alpha)max(|\mathbf{Y}(k)|^2 - E[|\mathbf{V}(k)|^2], 0), \quad (1)$$

where $|\hat{\hat{\boldsymbol{\theta}}}(k)|^2$ is the estimate of the kth frame's power spectrum at the output of the WPSS, α is the weighting for the previous frame, $|\hat{\boldsymbol{\theta}}(k-1)|^2$ is the power spectrum of the estimated signal of the previous frame, $|\mathbf{Y}(k)|^2$ is the power spectrum of the noisy signal, and $E[|\mathbf{V}(k)|^2]$ is the Power Spectral Density (PSD) of the noise. Here we use bold face letters to represent vectors. The WPSS then takes the square-root of the weighted power spectrum and combines it with the noisy phase to form its output $\hat{s}_{pss}(n,k)$. The LPC block uses the $\hat{s}_{pss}(n,k)$ to estimate the AR coefficients of the signal.

The WPSS procedure pre-processes the noisy signal so that the iteration starts at a point close to the maximum of the likelihood function, and is thus an initialization procedure. Initialization is crucial to EM approaches. A good initialization can make the convergence faster and prevent converging into a local maxima of the likelihood function. Several authors have suggested using an improved initial estimate of the parameters at the first iteration. In [4], Higher Order Statistics is used in the first estimation of AR parameters in order to improve the immunity to Gaussian noise. In [9], the noisy spectrum is first smoothed before the iteration begins. The initialization that is used here can be understood as using the likelihood maximum found in the previous frame as the starting point in the search of the maximum in the current frame, at the same time adapts to changes by incorporating new information from the PSS estimate. It can also be understood as a smoothed Power Spectral Subtraction method, noting the similarity between (1) and the Decision-Directed method used in [13]. Our experiments show that with this initialization procedure, an EM based approach can achieve faster convergence and higher SNR gain when the α is set appropriately.

Other authors have suggested sequential EM approaches in, e.g. [2] [3] [4] [5] [9]. These methods are sequential on a sample-to-sample basis. Thus the AR coefficients and the residual related parameters need to be estimated at every time instant. Our new algorithm is sequential frame-wise. This reduces computational complexity by exploiting the slow variation of the spectral envelopes (represented by the AR model). The system noise variance, on the other hand, needs a high temporal resolution estimation, and is discussed in the next section.

4 Kalman filtering with high temporal resolution signal model

Speech signals are known as non-stationary. Common practice is to segment the speech into short frames of 10 to 30 ms and assume a certain stationarity within the frame. Thus the temporal resolution of such a quasi-stationarity based processing equals the frame length. For voiced speech, the system noise usually exhibits large power variation within a frame (due to the impulse train structure), thus a much higher temporal resolution is desired. In this work, we allow the variance of the system noise to be indeed time variant. We estimate it by locally smoothing an estimate of the instantaneous power of the system noise.

4.1 The Kalman filtering solution

We use the following signal model,

$$\begin{aligned} s(n) &= \sum_{i=1}^{p} a_i s(n-i) + u(n) \\ y(n) &= s(n) + v(n) \end{aligned} \tag{2}$$

where the speech signal $s(n)$ is modeled as a pth-order AR process, and $y(n)$ is the observation, a_i is the ith AR parameter, the system noise $u(n)$ and the observation noise $v(n)$ are uncorrelated Gaussian processes. The system noise $u(n)$ models the excitation source of the speech signal and is assumed to have a time dependent variance $\sigma_u^2(n)$ that needs to be estimated. The observation noise variance σ_v^2 is assumed to change much slower, such that it can be seen as time invariant in the duration of interest and can be estimated from speech pause. In this work, we further assume that it is known. Equation (2) can be represented by the state space model

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{A}\mathbf{x}(n-1) + \mathbf{b}u(n) \\ y(n) &= \mathbf{h}\mathbf{x}(n) + v(n) \end{aligned} \tag{3}$$

where boldface letters represent vectors or matrices.

This is a standard state space model for the speech signal. Details about the state vector arrangement and the recursive solution equations are omitted here for brevity. Interested readers are referred to the classic paper [14]. We use the Kalman fixed-lag smoother in our experiment since it obtains the smoothing gain at the expense of delay only (again, see [14]. Though, note that in the proposed algorithm the system noise variance is truly time variant, whereas in the conventional Kalman filtering based speech enhancement the system noise variance is quasi-stationary).

4.2 Parameter estimation

The AR coefficients and the excitation variance should ideally be estimated jointly. However, this turns out to be a very complex problem. Here we also take an iterative approach. The AR coefficients are first estimated as described in Section 3, and then the excitation and its rapidly time-varying variance are estimated by the HTRM block, given the current estimate of the AR coefficients. The Kalman filter then uses the current estimate of the AR coefficients and the excitation variance to filter the noisy signal. The spectrum of the filtered signal is used in the next iteration to improve the estimate of the AR coefficients. It is again an approximation to the Maximum Likelihood estimation of the parameters, in which every iteration increases the conditional likelihood of the parameters and the signal.

The time-varying residual variance is estimated by the HTRM block. Given the AR coefficients, a Kalman filter takes the $\hat{s}_{pss}$ as input and estimate the system noise, which is essentially the linear prediction error of the clean signal. To distinguish this operation from the second Kalman filter, we call it the Prediction Error Kalman filter (PEKF). Instead of using a conventional linear prediction analysis to find the linear prediction error, we propose to use the PEKF because it has the capability to estimate the excitation source for the clean signal given an explicit model of noise in the observations. Noting that $\hat{s}_{pss}$ is the output of a smoothed Power Spectral Subtraction estimator, it contains both remaining noise and signal distortion. We model the joint contribution of the remaining noise and the signal distortion by a white Gaussian noise $z(n)$. The PEKF thus assumes the following state space model:

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{A}\mathbf{x}(n-1) + \mathbf{b}u(n) \\ \hat{s}_{pss}(n) &= \mathbf{h}\mathbf{x}(n) + z(n). \end{aligned} \tag{4}$$

Comparing with (3), the differences are: 1) now the $\hat{s}_{pss}$ becomes the observation, 2) the system noise $u(n)$ is now modeled as a Gaussian process with $constant$ variance within the frame, 3) the observation noise $z(n)$ has a smaller variance than $v(n)$ because the WPSS procedure has removed part of the noise power. The same Kalman solution as stated before is used to evaluate the prediction, $\hat{\mathbf{x}}(n|n-1)$, and the filtered estimation, $\hat{\mathbf{x}}(n|n)$. The prediction error is defined as $e(n) = \hat{\mathbf{x}}(n|n) - \hat{\mathbf{x}}(n|n-1)$. The reason that in the PEKF the system noise variance is modeled as constant within a frame is that we only use it as an initial estimate, and a finer estimate of the time variant variance is obtained at the output of the HTRM block. This is necessary since we can not use the estimate of the $\sigma_u^2(n)$ in the previous frame as the initialization, due to the fact that the proposed processing framework is not pitch-synchronous. We assume $z(n)$ to be zero-mean Gaussian with variance $\sigma_z^2 = \beta\sigma_v^2$, where β is a fractional scalar determined by experiments.

The high temporal resolution estimate of the system noise variance $\sigma_u^2(n)$ is obtained by local smoothing of the instantaneous power of $e(n)$. By a moving average smoothing using 2 or 3 points at each side of the current data point we get a quite good result. However, we found that a cubic spline smoothing yields better performance. The reason could be that the spline smoothing smoothes more in the valleys between two impulses than at the impulse peaks because of the large difference between the amplitudes of the impulse and the noise floor. This property of spline smoothing is desirable for our purpose since we want to maintain the dynamic range of the impulse as much as possible while smoothing out noise in the valleys. The cubic spline smoothing is implemented using the Matlab routine `csaps` with the smoothing parameter set to 0.1.

5 Experiments and results

We first define three objective quality measures used in this section, i.e., the signal to noise ratio (SNR), segmental SNR (segSNR), and Log-Spectral Distortion (LSD). The SNR is defined as the ratio of the total signal power to the total noise power in the utterance. SNR provides a simple error measure although its suitability for perceptual quality measure is questioned since it equally weights the frames with different energy while noise is known to be especially disturbing in low energy parts of the speech. We mainly use SNR as a convergence measure. Segmental SNR is defined as the average ratio of signal power to noise power per frame, and is regarded to be better correlated with perceptual quality than the SNR. The LSD is defined as the distance between two log-scaled DFT spectra averaged over all frequency bins [15]. We measure the LSD on voiced frames only. Common parameters are set as follows: the sampling frequency is 8 kHz, the AR model order is 10, the frame length is 160 samples. We aim at removing broad band noise from speech signals. In the experiments, the speech is contaminated by computer generated white Gaussian noise. The algorithm can be easily extended for the colored noise by augmenting the signal state vector and the transition matrix with the ones of the noise [8].

$Iter.$ \ α	**0.0**	**0.8**	**0.9**	**0.95**	**0.96**	**0.97**	**0.98**	**0.99**	**IEM**
1	9.45	10.39	10.86	11.22	11.31	**11.38**	**11.41**	**11.33**	10.36
2	10.57	11.07	**11.26**	**11.36**	**11.37**	11.37	11.33	11.21	11.06
3	10.94	**11.12**	11.20	11.22	11.22	11.20	11.17	11.06	**11.17**
4	**10.99**	11.06	11.09	11.09	11.08	11.07	11.05	10.97	11.11

Table 1: Output SNR of IEM+WPSS at different α and IEM.

We then compare the performance of the IEM with and without WPSS initialization, in order to show the effectiveness of the WPSS initialization. The two system

configurations are as in Fig. 2. When it is without the WPSS, the IEM is initialized by estimating the AR coefficients from the noisy signal. In the original IEM [2], the observation noise variance is estimated iteratively as part of the EM estimation and the system noise variance is obtained from the variance of the LPC residual. In this work, the observation noise variance is estimated from the speech pause. Utilizing this information, for the IEM, the initial estimate of the system noise variance is obtained by subtracting the noise variance from the LPC residual variance. We found that this modification improves the SNR gains by about 2 dB. In the sequel, we refer to the modified version as the IEM. Table 1 shows the output SNR of the IEM with WPSS initialization (IEM+WPSS) at different α and the IEM versus the number of iterations. The input signal is 3.6 seconds of male speech corrupted by white Gaussian noise at 5 dB SNR. By the SNR measure, the IEM converges at the third iteration. While for the IEM+WPSS, the iteration of convergence is dependent of α. When α is greater than 0.96, the algorithm achieves convergence at the first iteration. With α larger than 0.98 the SNR improvement decreases. Experiments on more speech samples and SNR levels show a consistent trend. Thus the α is decided to be 0.98. The result shows that the IEM with WPSS initialization ($\alpha = 0.98$) can achieve convergence at the first iteration and obtain even higher SNR gain than the IEM with three iterations.

Next, to determine the values of the weighting factor α and the remaining-noise-factor β for the proposed iterative Kalman filtering (IKF) algorithm, the algorithm is applied to 16 sentences from the TIMIT corpus added with white Gaussian noise at 5 dB SNR with various values of α and β. As is for the IEM+WPSS, the number of iterations needed for convergence of IKF is dependent of the parameters. The combination of α and β that makes convergence at the first iteration and gives the best result is chosen. By balancing the noise reduction and signal distortion, we choose the combination: $\alpha = 0.95, \beta = 0.5$.

It is observed in this experiment that for an α smaller than 0.98, setting β to a value larger than 0 results in a great improvement in the SNR, segSNR, and LSD, in comparison to when β is 0. Note that when β equals 0, the PEKF is reduced to the conventional linear prediction error filter. This suggests that the prediction-error Kalman filter succeeds in modeling and reducing the remaining noise in the excitation source that can not be modeled by the linear prediction error filter. When the α is larger than 0.98, setting β to a positive value does not improve the SNR and LSD, but still significantly improves the segSNR.

Now we compare the IKF with the base line IEM, and the IEM+WPSS algorithm. The results averaged on 30 TIMIT sentences (the training set used in the parameter selection is not included) are listed in Table 2. Significant improvement in all the three performance measures is observed, especially the segmental SNR. The only exception is the LSD at 0 dB. To confirm the subjective quality improvement, we apply a Degradation Mean Opinion Score (DMOS) test on the enhanced speech by the IKF and IEM,

with 10 untrained listeners. The result is shown in Tab 3. The listening test reveals that the background noise level in the IKF output is perceived to be significantly lower than the IEM. Besides, the low score of IEM is attributed to the annoying musical artifact, which is greatly reduced in the IKF. At input SNR higher than 15 dB, the background noise in the IKF enhanced speech is reduced to almost inaudible without introducing any major artifact.

Input	Methods	SNR[dB]	segSNR[dB]	LSD[dB]
20dB	**IKF**	23.13	12.60	1.89
	IEM+WPSS	22.75	11.42	2.08
	IEM	22.72	11.61	2.07
15dB	**IKF**	19.16	9.48	2.46
	IEM+WPSS	18.74	7.79	2.68
	IEM	18.69	8.13	2.65
10dB	**IKF**	15.37	6.65	3.15
	IEM+WPSS	14.96	4.36	3.33
	IEM	14.85	4.76	3.30
5dB	**IKF**	11.71	4.07	4.06
	IEM+WPSS	11.40	1.13	3.96
	IEM	11.18	1.56	3.97
0dB	**IKF**	8.25	1.81	5.24
	IEM+WPSS	8.11	-1.95	4.54
	IEM	7.81	-1.44	4.67

Table 2: Performance comparison. White Gaussian noise.

15dB	**IKF**	3.92	10dB	**IKF**	3.12	5dB	**IKF**	2.14
	IEM	2.25		**IEM**	1.98		**IEM**	1.64
	noisy	2.11		**noisy**	1.79		**noisy**	1.63

Table 3: DMOS scores.

6 Conclusion

In this paper, a new iterative Kalman filtering based speech enhancement scheme is presented. It is an approximation to the EM algorithm embracing the maximum likelihood principle. A high temporal resolution signal model is used to model voiced speech and the rapidly varying variance of the excitation source is estimated by a prediction-error Kalman filter. Distinct from other algorithms utilizing fine models for voiced speech, this approach avoids any voiced/unvoiced decision and pitch related parameter estimation. The convergence of the algorithm is obtained at the first iteration by introducing the WPSS initialization procedure. Performance evaluation shows significant improvements in three objective measures. Furthermore, informal listening indicates a significant reduction of musical noise. This result is confirmed by a DMOS subjective test.

References

[1] M. Feder, A. V. Oppenheim, and E. Weinstein, "Maximum likelihood noise cancellation using the EM algorithm," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 37, no.2, pp. 204–216, 1989.

[2] E. Weinstein, A. V. Oppenheim, and M. Feder, "Signal enhancement using single and multi-sensor measurements," *RLE Tech. Rep. 560, MIT, Cambridge, MA*, vol. 46, pp. 1–14, 1990.

[3] B. G. Lee, K. Y. Lee, and S. Ann, "An EM-based approach for parameter enhancement with an application to speech signals," *Signal Processing*, vol. 46, pp. 1–14, 1995.

[4] S. Gannot, "Algorithms for single microphone speech enhancement," *M.Sc. thesis, Tel-Aviv University*, Apr. 1995.

[5] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. on Speech and Audio*, vol. 6, pp. 373–385, July 1998.

[6] J. S. Lim and A. V. Oppenheim, "All-pole Modeling of Degraded Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASP-26, pp. 197–209, June 1978.

[7] J. H. L. Hansen and M. A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, 1991.

[8] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement," *IEEE Trans. on Signal Processing*, vol. 39, pp. 1732–1742, 1991.

[9] Z. Goh, K. Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. on Speech and Audio Processing*, vol. 7, No.5, pp. 510–524, 1999.

[10] C. Li and S. V. Andersen, "Inter-frequency Dependency in MMSE Speech Enhancement," *Proceedings of the 6th Nordic Signal Processing Symposium*, June 2004.

[11] ——, "A Block-based Linear MMSE noise reduction with a high temporal resolution modeling of the speech excitation," *EURASIP Journal on Applied Signal Processing*, vol. 18, pp. 2965–2978, 2005.

[12] ——, "Integrating Kalman filtering and multi-pulse coding for speech enhancement with a non-stationary model of the speech signal," *Proceedings of the 39th Asilomar Conference on Signals, Systems, and Computers*, June 2004.

[13] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, pp. 443–445, Apr. 1985.

[14] K. K. Paliwal and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering," *Proc.of ICASSP 1987*, vol. 12, pp. 177–180, Apr. 1987.

[15] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice Hall, 1988.

Paper E

Blind Identification of Non-Gaussian Autoregressive Models for Efficient Analysis of Speech Signals

Chunjian Li and Søren Vang Andersen

The paper has been published in
Proceedings, 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing.
May 14-19, 2006, Toulouse, France.

The layout has been revised.

Abstract

Speech signals, especially voiced speech, can be better modeled by non-Gaussian autoregressive (AR) models than by Gaussian ones. Non-Gaussian AR estimators are usually highly non-linear and computationally prohibitive. This paper presents an efficient algorithm that jointly estimates the AR parameters and the excitation statistics and dynamics of voiced speech signals. A model called the Hidden Markov-Autoregressive model (HMARM) is designed for this purpose. The HMARM models the excitation to the AR model using a Hidden Markov Model with two Gaussian states that have, respectively, a small and a large mean but identical variances. This formulation enables a computationally efficient exact EM algorithm to learn all parameters jointly, instead of resorting to pure numerical optimization or relaxed EM algorithms. The algorithm converges in typically 3 to 5 iterations. Experimental results show that the estimated AR parameters have much lower bias and variance than the conventional Least Squares solution. We also show that the new estimator has a very good shift-invariance property that is useful in many applications.

1 Introduction

Autoregressive (AR) modeling has been one of the most important techniques in speech signal processing. While the classical Least Squares (LS) solution, also known as LPC analysis, is computationally simple, it relies on a Gaussian AR model assumption. However, many important natural signals, including speech signals, are found to be far from Gaussian. The mismatch of a Gaussian model to a non-Gaussian signal causes an unnecessarily large variation in the estimates. This is supported by the fact that the Cramer-Rao bound for the variances of the AR estimators is lower in the non-Gaussian case than in the Gaussian case [1]. Smaller variances of AR estimators are desirable in many speech processing applications. As an example, in linear predictive coding, when a sustained vowel is segmented into overlapping frames that are subsequently encoded, small variance and shift-invariance property of the estimates of AR parameters are very beneficial in reducing the entropy and thus the needed bit rate for encoding the AR parameters. Non-Gaussian modeling of speech signals also reduces the bias of the AR estimator caused by the spectral sampling effect of the impulse train in voiced speech excitations. Applications in speech synthesis, speech recognition, and speech enhancement can benefit from these properties of non-Gaussian AR modeling.

We see the non-Gaussian AR model estimation problem as a blind system identification problem since the AR parameters and the non-Gaussian statistics of the excitation need to be estimated jointly. Reported works in this field include Higher Order Statistics (HOS) based methods (see [2] for a comprehensive review), Gaussian Mixture Model (GMM) based methods [1, 3, 4] and non-linear dynamical methods [5].

The HOS-based methods do not require explicit knowledge of the excitation probability density function (pdf), but tend to produce high-variance estimates when the length of the data record is small [3] and are associated with high computational complexity due to the bispectrum calculation. The GMM-based methods estimate their parameters using the Maximum Likelihood (ML) criterion. Since the exact ML solution for non-Gaussian signals typically involves solving a set of highly non-linear equations, it has to be solved by computationally complex numerical algorithms, or by solving for an approximation of the ML solution. In [1], the ML solution is solved by a conventional Newton-Raphson optimization algorithm. In [4], the AR parameters and the excitation probability density function (pdf) are separately estimated in a recursive manner to approximate the joint estimation in a tractable way. In [3], the AR parameters and the excitation pdf are estimated by a generalized EM (GEM) algorithm, which relaxes from the standard EM algorithm by breaking the multi-dimensional optimization into recursive one-dimensional optimizations. The price to pay for the GEM is a slower convergence rate than the EM. The non-linear dynamic method proposed in [5] estimates the coefficients of an inverse filter by minimizing a dynamic-based complexity measure called phase space volume (PSV). This method does not assume any structure of the excitation, but the computation of PSV is rather involved.

Most of the reported non-Gaussian AR modeling techniques are for general purposes. While being applicable to any probability distribution, this also makes them less efficient in handling speech signals, whose production mechanism is well known and implies powerful structures in the signal. In this paper, we propose an algorithm that is designed to exploit the structure of voiced speech signals, aiming at better computational efficiency and data efficiency. The algorithm jointly estimates the AR parameters and the excitation statistics and dynamics based on a ML criterion. Here the voiced speech signal is modeled by a Hidden Markov-Autoregressive Model (HMARM), where the excitation sequence is modeled by a Hidden Markov Model (HMM) that has two states with Gaussian emission densities of different means but same variances and then convolved with an AR filter. The HMARM parameters can be learned efficiently by an exact EM algorithm consisting of a set of linear equations. This model is different from the Linear Predictive HMM (LP-HMM), or Autoregressive HMM (AR-HMM) used in [6] and [7]. The AR-HMM applies its dynamic modeling on tracking the AR model variation along frames, while the proposed HMARM applies dynamic modeling on tracking the impulse train structure of the excitation within a frame.

The remainder of this paper is organized as follows. Section 2 describes the problem formulation and derives the EM algorithm. The algorithm is evaluated with synthetic signals and speech signals in Section 3. Conclusion is made in Section 4.

2 The Method

The speech production mechanism is well modeled by the excitation-filter model, where an AR(p) filter models the vocal tract resonance property and an impulse train models the excitation of voiced speech. To improve naturalness of the speech, a white noise component is added to the impulse train. This can be expressed in the following equations:

$$x(t) = \sum_{k=1}^{p} g(k)x(t-k) + r(t) \tag{1}$$

$$r(t) = v(t) + u(t), \tag{2}$$

where $x(t)$ is the signal, $g(k)$ is the kth AR coefficient, and $r(t)$ is the excitation. The excitation sequence is the sum of an impulse train $v(t)$ and a white Gaussian noise sequence $u(t)$ with zero mean and variance σ^2. This noisy impulse train structure is perfectly suitable for stochastic dynamic modeling. We design a two-state HMARM whose diagram is shown in Fig.1. The state q_t at time t selects according to the state transition probability $a_{q_{t-1}q_t}$ one of two states. The emission pdfs of the two states are Gaussian pdfs with identical variances σ^2, and a small mean $m_r(1)$ and a large mean $m_r(2)$ respectively. The small mean is close to zero, and the large mean is equal to the amplitude of the impulses. The emission outcome constitutes the excitation sequence $r(t)$, which is independent of $r(l)$ for $l \neq t$ and only dependent on the state q_t. The excitation $r(t)$ is then convolved with an AR(p) filter with coefficients $[g(1), \cdots, g(p)]$ to produce the observation signal $x(t)$. The objective of the algorithm is to learn the model parameters $\boldsymbol{\phi} = [\mathbf{A}, m_r(1), m_r(2), \sigma^2, g(1), \cdots, g(p)]$ given a frame of signal $\mathbf{x}$ with length T, where the state transition matrix $\mathbf{A} = (a_{ij})$, with $i, j \in (1, 2)$.

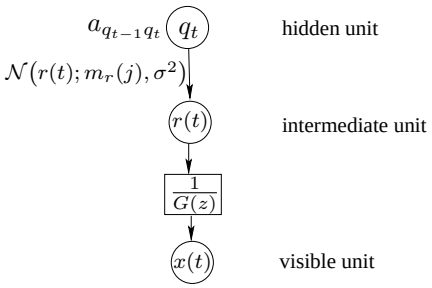


Figure 1: A generative data structure of the HMARM.

We now define the notations for the HMARM model. Let $\alpha(j,t)$ and $\beta(i,t)$ denote the forward and backward likelihoods as defined in the standard HMM [8], a_{ij}

denote the state transition (state i to state j) probability, $b_r(j,t)$ denote the observation pdf (emission pdf) of the excitation $r(t)$ given the state $q_t = j$, which is a Gaussian distribution

$$b_r(j,t) = \mathcal{N}\big(r(t); m_r(j), \sigma^2\big), \tag{3}$$

and $b_x(j,t)$ denote the observation pdf of the signal $x(t)$ given the state $q_t = j$. From (1) and (3), $b_x(j,t)$ can be shown to be a Gaussian process with a varying mean $m_x(j,t)$,

$$b_x(j,t) = \mathcal{N}\big(x(t); m_x(j,t), \sigma^2\big), \tag{4}$$

where

$$m_x(j,t) = \sum_{k=1}^{p} g(k)x(t-k) + m_r(j). \tag{5}$$

The forward and backward likelihood inductions are given by

$$\alpha(j,t) = \left[\sum_{i=1}^{N} \alpha(i,t-1)a_{ij}\right] b_x(j,t), \tag{6}$$

$$\beta(i,t) = \left[\sum_{j=1}^{N} a_{ij} b_x(j,t+1)\beta(j,t+1)\right], \tag{7}$$

respectively. Now define $\xi(i,j,t)$ to be the probability of being in state i at time t and in state j at time $t+1$, i.e. $\xi(i,j,t) = p(q_t = i, q_{t+1} = j|\mathbf{x}, \boldsymbol{\phi})$. One can evaluate $\xi(i,j,t)$ by

$$\xi(i,j,t) = \frac{\alpha(i,t)a_{ij}b_x(j,t+1)\beta(j,t+1)}{\sum_{t=0}^{T-1} a_{q_t q_{t+1}} b_x(q_{t+1}, t+1)}. \tag{8}$$

Define $\gamma(i,t) = \sum_{j=1}^{N} \xi(i,j,t)$. It can then be shown that the quantity $\sum_{t=1}^{T-1} \gamma(i,t)$ represents the expected number of transitions made from state i, and $\sum_{t=1}^{T-1} \xi(i,j,t)$ represents the expected number of transitions from state i to state j [8].

Now we derive the EM algorithm. Let bold face letters $\mathbf{x}$ and $\mathbf{q}$ denote a frame of signal and the state vector of the corresponding frame of excitation, respectively. We define the complete data to be $(\mathbf{x}, \mathbf{q})$. Instead of maximizing the log-likelihood $\log p(\mathbf{x}|\boldsymbol{\phi})$ directly, we maximize the expectation of the complete data likelihood $\log p(\mathbf{x}, \mathbf{q}|\boldsymbol{\phi})$ over the states $\mathbf{q}$ given the data $\mathbf{x}$ and current estimate of $\boldsymbol{\phi}$, denoted by $\tilde{\boldsymbol{\phi}}$. So the

function to be maximized in each iteration is written as:

$$Q(\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}) = \sum_{\mathbf{q}} \frac{p(\mathbf{x}, \mathbf{q}|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \log p(\mathbf{x}, \mathbf{q}|\boldsymbol{\phi}) \tag{9}$$

$$= \sum_{\mathbf{q}} \frac{p(\mathbf{x}, \mathbf{q}|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \Big(\sum_{t=1}^{T} \log a_{q_{t-1}q_t} + \sum_{t=1}^{T} \log b_x\big(q_t, x(t)\big) \Big) \tag{10}$$

$$= \sum_i \sum_j \sum_t \frac{p(\mathbf{x}, q_{t-1} = i, q_t = j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \log a_{q_{t-1}q_t} + \sum_j \sum_t \frac{p(\mathbf{x}, q_t = j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \log b_x\big(q_t, x(t)\big), \tag{11}$$

where (10) follows from the identity

$$p(\mathbf{x}, \mathbf{q}|\boldsymbol{\phi}) = \prod_{t=1}^{T} a_{q_{t-1}q_t} b_x\big(q_t, x(t)\big),$$

and (11) follows from the first order Markov assumption. The first term in (11) concerns only a_{ij} and the second term concerns the rest of the parameters. Thus the optimization can be done on the two terms separately. The re-estimation equation of a_{ij} is found by the Lagrange multiplier method, and is identical to the standard Baum-Welch re-estimation algorithm:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} p(\mathbf{x}, q_{t-1} = i, q_t = j|\tilde{\boldsymbol{\phi}})}{\sum_{t=1}^{T-1} p(\mathbf{x}, q_{t-1} = i|\tilde{\boldsymbol{\phi}})} = \frac{\sum_{t=1}^{T-1} \xi(i, j, t)}{\sum_{t=1}^{T-1} \gamma(i, t)}. \tag{12}$$

We denote the second term of (11) by $Q(\boldsymbol{\phi}, \hat{b})$. Following (1) and (4) we can write

$$Q(\boldsymbol{\phi}, \hat{b}) = \sum_j \sum_{t=1}^{T-1} \frac{p(\mathbf{x}, q_t = j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \Big(\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \Big(x(t) - m_x(j, t) \Big)^2 \Big). \tag{13}$$

The re-estimation equations of the rest of the parameters are found by setting the partial derivatives of (13) to zero, and solving the equation system. For $g(k)$, we have p

equations:

$$\sum_j \sum_{t=1}^{T-1} \gamma(j,t)\Big(x(t) - m_x(j,t)\Big)x(t-k) = 0, \quad k = 1, \cdots, p. \tag{14}$$

where $\gamma(j,t) = \frac{p(\mathbf{x}, q_t = j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})}$ is now interpreted as the posterior of state j at time t given the observation $\mathbf{x}$ and $\tilde{\boldsymbol{\phi}}$. For $m_r(j)$, we get two equations:

$$\sum_t^{T-1} \gamma(j,t)\Big(x(t) - m_x(j,t)\Big) = 0, \quad j = 1, 2. \tag{15}$$

For σ^2, we get

$$\widehat{\sigma^2} = \frac{\sum_j \sum_t^{T-1} \gamma(j,t)\Big(x(t) - m_x(j,t)\Big)^2}{\sum_j \sum_{t=1}^{T-1} \gamma(j,t)}. \tag{16}$$

Equation (14) and (15) form $p+2$ coupled linear equations which can be solved analytically. Then (16) can be solved by inserting the estimated $g(k)$ and $m_r(j)$.

In this model, $m_x(j,t)$ can be interpreted as the linear prediction of $x(t)$ taking into account the excitation dynamics, as shown in (5). The re-estimation equations also have intuitive interpretations. In (12), a_{ij} equals the expected number of transitions from state i to state j divided by the expected number of transitions made from state i; Equation (14) is a multi-state version of the orthogonality principle; Equation (15) tells that the prediction error weighted by state posterior is of zero mean; and (16) calculates the mean of the prediction error power weighted by the state posterior as the variance of the stochastic element of the signal.

The existence of linear solutions to the maximization of the Q function makes fast convergence. This is a direct benefit from our proposed signal model. Compared to the GMM-based method in [3], which has no analytical solution to the maximization of Q function, the HMM in our model is constrained to have states with identical emission variance. It is this constraint that renders the set of non-linear equations linear, without compromising the validity of the model.

A GMM with similar constraint can be used in place of the HMM in our signal model, and the EM equations can be derived in the same way as shown above with proper changes in the definition of α and β (and $\xi(i,j,t)$ is not needed in the GMM). In our experience, this constrained GMM-AR model results in a slower convergence rate and slightly worse estimation accuracy than the HMARM. This is expected since the GMM lacks capability of dynamic modeling, while the impulse train does show a clear dynamic structure.

Finally, we point out an implementation issue of the HMARM estimation. Since

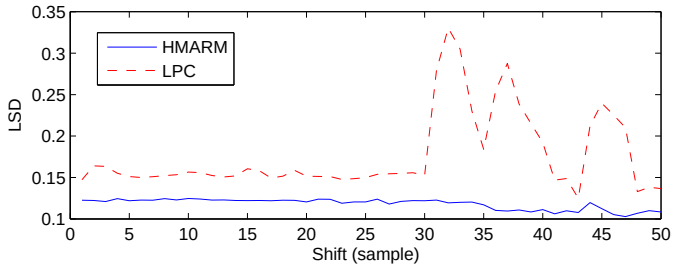
the signal model is a causal dynamic model and the analysis is usually frame-based, the ringing from the last impulse of the previous frame has an undesired impact on the current frame estimates. This is because the estimator does not see the previous impulse but its effect is there. This could sometimes degrade the performance mildly. We therefore suggest to do a pre-processing that removes the ringing from the previous frame, or simply set the signal before the first impulse to zeros. The latter is used in our experiments.

3 Experimental results

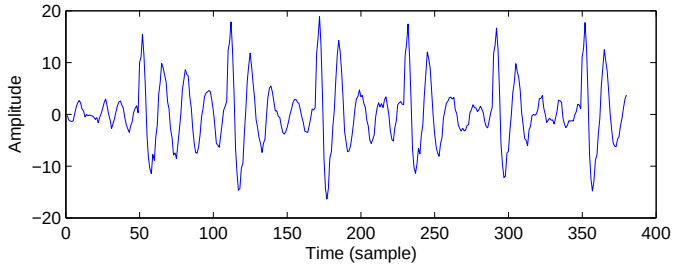
We now experimentally compare the spectral distortion, the variance, and the bias of the AR parameters estimated by the proposed HMARM analysis and the LPC analysis. To get different realizations of an AR process, we shift a rectangular window along a long segment of the signal by one sample each time. Every shift produces a different realization frame of the AR process. A small variance of the estimates based on shifted realizations is also known as the shift-invariance property. The LPC analysis has a poor shift-invariance property when it is applied to voiced speech. This is because its underlying Gaussian model does not fit the non-Gaussian nature of the excitation of the voiced speech.

First, to have access to the true values of the AR parameters of a signal, we use a synthetic signal that mimics a voiced speech signal. The signal is analyzed by the HMARM and the LPC analysis respectively for 50 realizations with a frame length of 320 samples. The 50 realizations of estimated AR spectra are compared to the true AR parameters and the difference is measured by the Log-Spectral Distortion (LSD) measure. The LSD versus the shift is shown in Fig 2. It is clear that the proposed method has a flat distortion surface and this surface is lower than the LPC's. It is important to note that the LPC analysis encounters huge deviation from the true values in the second half of the plot. This is where a large "hump" in the signal comes into the analysis frame. The large humps in the signal are caused by the impulses in the excitation, which represent the non-Gaussian structure of the signal. The bias is 0.092 for the HMARM analysis, and compared to the 0.197 for the LPC analysis, accounts for an improvement of more than 6 dB. The variance is 0.128 for the HMARM and 9.69 for the LPC analysis, representing a variance reduction of 18 dB.

Second, we test the shift-invariance property with true speech signals. The AR spectra of four different sustained voiced phonemes are estimated 50 times with one sample shift each time. The frame length is set to 256 samples. The spectra are plotted in Fig 3. The estimates by the HMARM show good consistency, while the LPC analysis appears to be poor. In Fig. 4 we show the prediction residuals of the signal using the AR parameters estimated by the HMARM and the LPC respectively. It is clear that



(a)



(b)

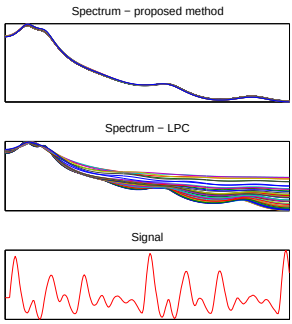
Figure 2: (a): The Log-Spectral Distortion of the AR spectra. (b): the synthetic signal waveform used in the test.

the residual of the HMARM has more prominent impulses, and less correlation in the valleys. From, as one example, a speech coding point of view, the lower variance of the AR estimates reduces the entropy of the AR parameters, and the more impulsive residual is also easier to code.

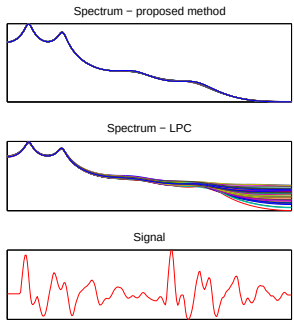
As it is well known that a properly chosen window can reduce the variance of the LPC estimates, we also conducted comparisons between the HMARM analysis and the Hamming-windowed LPC analysis. For the synthetic signal, the variance of the Hamming-windowed LPC is 1.197, which is still 9.7 dB higher than that of the HMARM. Although its variance is reduced, the Hamming-windowed LPC in general suffers from larger bias and lower spectral resolution. Due to space limit, more results will be presented in a following paper.

4 Conclusion

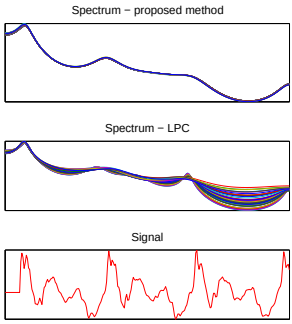
A non-Gaussian AR model is proposed to model the voiced speech signal. This model enables an efficient EM algorithm that consists of a set of linear equations. The algorithm jointly estimates the AR parameters of the signal and the dynamics of the exci-



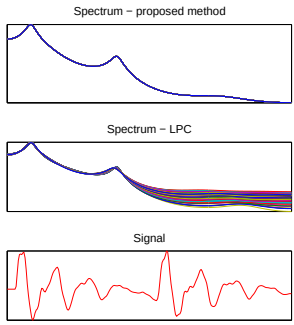
(a) Vowel /o/ in “short”.



(b) Vowel /e/ in “red”.



(c) Vowel /a/ in “began”.



(d) Vowel /ou/ in “shoulder”

Figure 3: The AR spectra estimated by HMARM and LPC analysis.

tation that is highly non-Gaussian in the voiced speech case. The experimental results using synthetic signals and real speech signals show that the algorithm has a good shift-invariance property, and the variance and bias are significantly smaller than the classical LPC analysis.

References

[1] D. Sengupta and S. Kay, “Efficient estimation of parameters for non-Gaussian autoregressive processes,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37. No.6, pp. 785–794, 1989.

[2] C. L. Nikias and M. R. Raghuveer, “Bispectrum estimation: a digital signal processing framework,” *Proc. IEEE*, vol. 75, pp. 869–891, 1987.

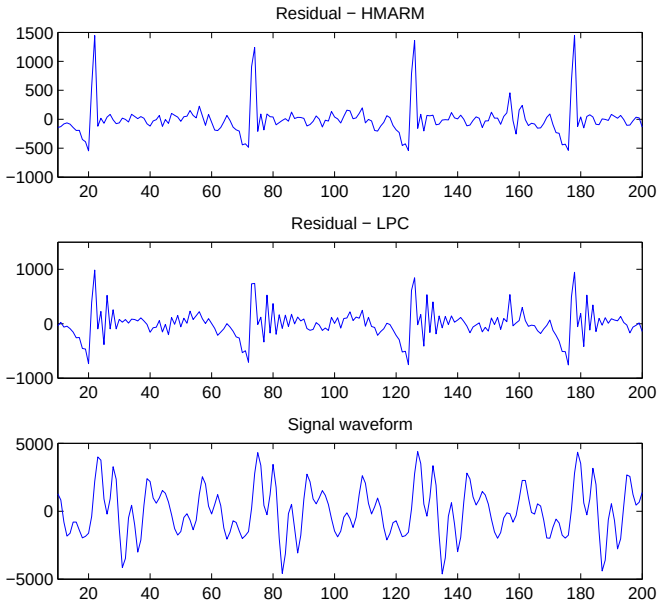


Figure 4: Prediction residuals by the HMARM and the LPC analysis.

[3] S. M. Verbout, J. M. Ooi, J. T. Ludwig, and A. V. Oppenheim, "Parameter estimation for autoregressive Gaussian-Mixture processes: the EMAX algorithm," *IEEE Trans. on Signal Processing*, vol. 46. No.10, pp. 2744–2756, 1998.

[4] Y. Zhao, X. Zhuang, and S.-J. Ting, "Gaussian mixture density modeling of non-Gaussian source for autoregressive process," *IEEE Trans. on Signal Processing*, vol. 43. No.4, pp. 894–903, 1995.

[5] H. Leung, S. Wang, and A. M. Chan, "Blind identification of an autoregressive system using a non-linear dynamical approach," *IEEE Trans. on Signal Processing*, vol. 48. No.11, pp. 3017–3027, 2000.

[6] A. Poritz, "Linear predictive hidden Markov models and the speech signal," *ICASSP'82*, vol. 7, pp. 1291–1294, 1982.

[7] B.-H. Juang and L. R. Rabiner, "Mixture autoregressive Hidden Markov Models for speech signals," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-33. No.6, pp. 1404–1413, 1985.

[8] L. R. Rabiner and B. H. Juang, "An introduction to Hidden Markov Model," *IEEE ASSP Magazine*, pp. 4–16, Jan. 1986.

Paper F

Efficient Blind System Identification of Non-Gaussian Auto-Regressive Models with Dynamic Modeling

Chunjian Li and Søren Vang Andersen

The paper has been accepted for publication in
IEEE Transactions on Signal Processing, 2006

The layout has been revised.

Abstract

We have previously proposed a blind system identification method that exploits the underlying dynamics of non-Gaussian signals in [1]. The signal model being identified is an Auto-Regressive (AR) model driven by a discrete-state Hidden Markov process. An exact EM algorithm was derived for the joint estimation of the AR parameters and the HMM parameters. In this paper, we extend the system model by introducing an additive measurement noise. The identification of the extended system model becomes much more complicated since the system output is now hidden. We propose an exact EM algorithm that incorporates a novel Switching Kalman Smoother, which obtains optimum nonlinear MMSE estimates of the system output based on the state information given by the HMM filter. The exact EM algorithms for both models are obtainable only by appropriate constraints in the model design, and have better convergence properties than algorithms employing generalized EM algorithm or empirical iterative schemes. The proposed methods also enjoy good data efficiency since only second order statistics is involved in the computation. The signal models are general and suitable to numerous important signals, such as speech signals and base-band communication signals. This paper describes the two system identification algorithms in an integrated form, and provides supplementary results to the noise-free model and new results to the extended model with applications in speech analysis and channel equalization.

1 Introduction

One of the recent trends in signal processing is to exploit non-Gaussianity or non-stationarity of the signals to accomplish tasks that are generally impossible for traditional linear estimators, e.g., blind source separation, blind channel equalization, and blind system identification. Blind system identification (BSI) solves the fundamental problem residing in most signal processing fields: estimating the system parameters from system output only. In this definition of BSI, the model selection is a preliminary step to the actual identification process. Model selection is usually done according to prior knowledge of the underlying physics of the system. So the task of the BSI is to extract *a posteriori* information from the system output. A good model selection should facilitate the identification process without compromising the validity of the model much.

In this work, we present two signal models that have efficient identification solutions. On one hand, they are general enough to accommodate many important signals such as speech signals and base band communications signals with the presence of Inter-Symbol Interference (ISI). On the other hand, the efficiency of the algorithms comes from the prior knowledge of the specific signal structure carried by the model.

The first system model consists of a linear time-invariant AR filter excited by a

first-order discrete-state Hidden Markov process. In the speech analysis application, the AR filter models the resonant property of the vocal tract, and a two-state Hidden Markov process models the excitation to the filter as a noisy impulse train. The task of system identification here is to jointly estimate the AR coefficients and the excitation dynamics, which contains information about the impulse position, the impulse amplitude, and the noise variance, under a certain optimum criterion. By the joint estimation, the highly non-Gaussian impulse train structure of the excitation no longer affects the AR estimation as it does in the classic Least Squares (LS) solution. The LS methods, such as the auto-correlation method, a.k.s. the LPC analysis, assumes a Gaussian signal model. The consequence of the mismatch of Gaussian model to non-Gaussian signals is an unnecessarily large variation in the estimates. This is supported by the fact that the Cramer-Rao bound for the variances of the AR estimators is lower in the non-Gaussian case than in the Gaussian case [2]. Estimating the AR parameters taking into account the impulse structure of the excitation can also reduce bias. This bias is present in the LPC analysis because of the spectral sampling effect of the impulse train. We will show that the AR spectra estimated by our method have smaller variance and bias and a better shift invariance property than the LPC analysis. These properties are useful in a wide range of speech processing fields, such as speech coding, pitch modification, speech recognition, and speech synthesis. The identification is done through an exact EM algorithm that consists of forward-backward calculations of state posterior and solving a small linear equation system iteratively. Initialized with the LPC estimates, using only a few dozens of samples, the algorithm converges in typically 3 to 5 iterations.

Application of this model to the blind channel equalization problem is also demonstrated in this paper. To combat ISI in a dispersive channel, channel equalizers are used in many communication systems before decoding the signal. When neither the channel response nor the transmitted-symbol statistics are known *a priori*, hence the name blind equalization, the channel response and transmitted symbols need to be estimated jointly. Most established blind equalization methods presume the channel to be FIR. Our blind equalization method, instead, is based on an assumption of an IIR all-pole channel model with the following arguments: 1) The use of an AR channel model can reduce the computational complexity dramatically by exploiting the Markovian property of the channel; 2) In channels that exhibit resonance property, such as wireline channels, an AR model is probably more realistic than an FIR model; 3) An AR model with a sufficiently high order can approximate any ARMA or MA model very well. To be specific, the AR filter models the channel response, and the Hidden Markov process models the sampled base-band signals. The algorithm exploits the underlying dynamics and non-Gaussianity of the finite alphabet symbol sequence to accomplish system identification. An example of equalizing an MA channel is also demonstrated.

In the second system model, observation noise is taken into account. Now, the model consists of a linear time-invariant AR filter excited by a first-order discrete-state Hidden

Markov process, and the measurements of the system output are perturbed by white Gaussian noise. The identification algorithm must jointly estimate the AR parameters, the excitation dynamics, and the measurement noise variance. The introduction of measurement noise complicates the problem significantly. This is because that the simplicity of the first algorithm partly comes from the fact that the AR model aggregates the state information in the most recent system output samples, which are not directly observable now due to the presence of measurement noise. We adopted a layered data structure with Markov property between layers, which is analogous to the one used in the Independent Factor Analysis [3]. The EM algorithm thus involves a nonlinear MMSE smoother, which provides estimates of the conditional first and second moments of the system output needed in the parameter estimations. We propose a nonlinear MMSE smoother that can be seen as a variant of the soft-decision Switching Kalman Filter [4], where the states control the discrete inputs to the AR filter, and the switching relies on the *a posteriori* probability of states estimated by a forward-backward algorithm. The EM algorithm thus iterates between the nonlinear MMSE smoothing and the ML parameter estimations.

The introduction of measurement noise modeling in the second system model is a major extension to the first system model. The second method is thus noise robust and applicable in adverse environments, although with a price of higher computational complexity. In its application to robust spectrum estimation of speech signals, the algorithm gives better estimates of the signal spectra than reference methods do, under moderate noise conditions. Established iterative estimators based on Gaussian AR models are known to have convergence problems, thus an empirical termination is required [5] [6]. They also require prior knowledge of measurement noise statistics. The proposed algorithm does not require prior knowledge of the noise statistics, and its convergence is guaranteed. Applications to channel equalization under moderate noise conditions are also demonstrated. Simulations show that the proposed algorithm has better estimates of the channel response and the transmitted symbols than the Least Squares method.

The remainder of the paper is organized in the following way: Section 2 introduces the two signal models and derives the EM algorithms for blind system identification. In Section 3 the proposed algorithms are applied to solving problems in speech analysis, noise robust spectrum estimation, and blind channel equalizations with and without measurement noise. We conclude in Section 4.

2 Method

We consider the stochastic source-filter model, in which a linear time invariant (LTI) filter is excited by a stochastic process with a certain statistic property. When the excitation is stationary and Gaussian, the Least Squares method provides an optimum

solution to the system identification problem. Nevertheless, many important signals are far from Gaussian. Voiced speech signals and modulated communication signals transmitted through a dispersive channel are just two examples of such signals. A common characteristic of the above mentioned two non-Gaussian signals is that the excitation can be viewed as a sequence of symbols drawn from a finite alphabet, with possibly additive noise. More specifically, for voiced speech, the excitation is well modeled by an impulse train with additive white Gaussian noise [7]. This noisy impulse train structure can be characterized by a two-state symbol sequence. While an M-ary Pulse Amplitude Modulation (PAM) signal can be characterized by an M-state symbol sequence. The probability distribution functions (pdfs) of these discrete state excitations are thus multi-modal, and possibly asymmetric (as is for the impulse train). Based on this observation, either a Gaussian Mixture Model (GMM) or a Hidden Markov Model (HMM) with discrete states is suitable to characterize the statistics of such excitations. When such non-Gaussian excitations are filtered by an AR filter, we term the system model a Hidden Markov-Auto Regressive Model (HMARM) or a Gaussian Mixture-Auto Regressive model (GMARM), respectively. We will show in the following sections that when the emission pdfs of all states are constrained to be Gaussian pdfs with identical variance, both the HMARM and the GMARM have exact EM algorithms for their identifications. Whereas, the HMM is preferable in modeling the excitation because of its capability of modeling the underlying temporal structure that is not captured by the GMM, which is still a static statistical model. Therefore, the following presentation will mainly focus on the HMARM with a brief discussion on the advantage of the HMM over the GMM in modeling temporal structure.

In Section 2.1, we present the HMARM and its identification without measurement noise. Section 2.2 deals with the identification of HMARM with its output perturbed by white Gaussian noise, which is termed the Extended-HMARM.

2.1 The HMARM and its identification

For an AR(p) filter excited by a Hidden Markov sequence, we have the following system model:

$$x(t) = \sum_{k=1}^{p} g(k)x(t-k) + r(t) \tag{1}$$

$$r(t) = v(t) + u(t), \tag{2}$$

where $x(t)$ is the observed signal (system output), $g(k)$ is the kth AR coefficient, and $r(t)$ is the excitation. The excitation is a Hidden Markov process, i.e., a first order Markov chain $v(t)$ plus white Gaussian noise $u(t)$ with zero mean and variance σ^2. A diagram of the data structure of the HMARM is shown in Fig. 1, which adopts a

layered data structure analogous to the one used in [3]. The state q_t at time t selects according to the state transition probability $a_{q_{t-1}q_t}$ one of M states. The emission pdfs of the states are Gaussian pdfs with the *same variance* σ^2, and means $m_r(j)$, $j \in (1, \cdots, M)$, respectively. The emission outcome constitutes the excitation sequence $r(t)$, which is independent of $r(l)$ for $l \neq t$ and only dependent on the state q_t. The excitation $r(t)$ is then convolved with an AR(p) filter with coefficients $[g(1), \cdots, g(p)]$ to produce the observation $x(t)$. The objective of the identification algorithm is to learn the model parameters $\boldsymbol{\phi} = [\mathbf{A}, m_r(1), \cdots, m_r(M), \sigma^2, g(1), \cdots, g(p)]$ given a frame of signal with length T, where the state transition matrix is denoted by $\mathbf{A} = (a_{ij})$, $i, j \in (1, \cdots, M)$.

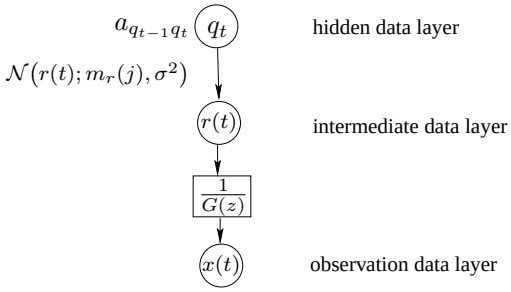


Figure 1: A generative data structure of the HMARM.

We now define some HMM type notations. Let $\alpha(j,t)$ and $\beta(i,t)$ denote the forward and backward likelihoods as defined in [8], and a_{ij} denote the state transition probability (from state $q_t = i$ to state $q_{t+1} = j$), and $b_r(j,t)$ denote the emission pdf of state $q_t = j$ observed at the intermediate layer $r(t)$. Follows from (2), the emission pdf $b_r(j,t)$ takes on a Gaussian distribution

$$b_r(j,t) = \mathcal{N}\big(r(t); m_r(j), \sigma^2\big). \tag{3}$$

Now, let $b_x(j,t)$ denote the emission pdf of state $q_t = j$ observed at the observation data layer $x(t)$. It is difficult to deduce this pdf from top layer down to the bottom layer because of the filtering. But we can use the autoregressive property of the filter, i.e., the p most recent system outputs and the current input state define the current output uniquely. From (1), (2) and (3), $b_x(j,t)$ can be shown to be a Gaussian pdf with a *time varying* mean $m_x(j,t)$,

$$b_x(j,t) = \mathcal{N}\Big(x(t); m_x(j,t), \sigma^2\Big), \tag{4}$$

where

$$m_x(j,t) = \sum_{k=1}^{p} g(k)x(t-k) + m_r(j). \tag{5}$$

The forward and backward likelihood inductions are given by

$$\alpha(j,t) = \left[\sum_{i=1}^{M} \alpha(i,t-1)a_{ij}\right] b_x(j,t), \tag{6}$$

$$\beta(i,t) = \left[\sum_{j=1}^{M} a_{ij} b_x(j,t+1)\beta(j,t+1)\right], \tag{7}$$

respectively. Now define $\xi(i,j,t)$ to be the probability of being in state i at time t and in state j at time $t+1$, i.e. $\xi(i,j,t) = p(q_t = i, q_{t+1} = j|\mathbf{x}, \boldsymbol{\phi})$. One can evaluate $\xi(i,j,t)$ by

$$\xi(i,j,t) = \frac{\alpha(i,t)a_{ij}b_x(j,t+1)\beta(j,t+1)}{\sum_{t=1}^{T-1} a_{q_t q_{t+1}} b_x(q_{t+1}, t+1)}, \quad t \in [1, T-1]. \tag{8}$$

Define $\gamma(i,t) = \sum_{j=1}^{M} \xi(i,j,t)$. It can then be shown that the quantity $\sum_{t=1}^{T-1} \gamma(i,t)$ represents the expected number of transitions made from state i, and $\sum_{t=1}^{T-1} \xi(i,j,t)$ represents the expected number of transitions from state i to state j [8].

Now we are ready to derive the EM algorithm for identification. Let bold face letters $\mathbf{x}$ and $\mathbf{q}$ denote a frame of signal and the state vector of the corresponding frame of excitation, respectively. We define the complete data to be the concatenation of the observation data and the hidden data $(\mathbf{x}, \mathbf{q})$, as indicated in Fig. 1. The excitation $r(t)$ can not be treated as hidden data because once the parameters $\boldsymbol{\phi}$ are known, $r(t)$ is linearly dependent on the observation data. Hence we term it the intermediate data. Following the EM paradigm [9], we maximize, instead of the log-likelihood $\log p(\mathbf{x}|\boldsymbol{\phi})$ directly, the expectation of the complete data likelihood $\log p(\mathbf{x}, \mathbf{q}|\boldsymbol{\phi})$ over the states $\mathbf{q}$ given the observation $\mathbf{x}$ and current estimate of $\boldsymbol{\phi}$, which is denoted by $\tilde{\boldsymbol{\phi}}$. So the

function to be maximized in each iteration is written as [1]:

$$Q(\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}) = \sum_{\mathbf{q}} p(\mathbf{q}|\mathbf{x}, \tilde{\boldsymbol{\phi}}) \log p(\mathbf{x}, \mathbf{q}|\boldsymbol{\phi}) \tag{9}$$

$$= \sum_{\mathbf{q}} \frac{p(\mathbf{x}, \mathbf{q}|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \log p(\mathbf{x}, \mathbf{q}|\boldsymbol{\phi}) \tag{10}$$

$$= \sum_{\mathbf{q}} \frac{p(\mathbf{x}, \mathbf{q}|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \left(\sum_t \log a_{q_{t-1} q_t} + \sum_t \log b_x\big(q_t, x(t)\big) \right) \tag{11}$$

$$= \sum_i \sum_j \sum_t \frac{p(\mathbf{x}, q_{t-1} = i, q_t = j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \log a_{q_{t-1} q_t} + \sum_j \sum_t \frac{p(\mathbf{x}, q_t = j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \log b_x\big(q_t, x(t)\big), \tag{12}$$

where (11) follows from the identity

$$p(\mathbf{x}, \mathbf{q}|\boldsymbol{\phi}) = \prod_{t=1}^{T} a_{q_{t-1} q_t} b_x\big(q_t, x(t)\big),$$

and (12) follows from the first order Markov assumption. The first term in (12) concerns only a_{ij} and the second term concerns the rest of the parameters. Thus the optimization can be done on the two terms separately. The re-estimation equation of a_{ij} is found by the Lagrange multiplier method, and is identical to the standard Baum-Welch re-estimation algorithm [10]:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} p(\mathbf{x}, q_{t-1} = i, q_t = j|\tilde{\boldsymbol{\phi}})}{\sum_{t=1}^{T-1} p(\mathbf{x}, q_{t-1} = i|\tilde{\boldsymbol{\phi}})} = \frac{\sum_{t=1}^{T-1} \xi(i, j, t)}{\sum_{t=1}^{T-1} \gamma(i, t)}. \tag{13}$$

We denote the second term of (12) by $Q(\boldsymbol{\phi}, \hat{b})$. Following (1) and (4) we can write

$$Q(\boldsymbol{\phi}, \hat{b}) = \sum_j \sum_{t=1}^{T-1} \frac{p(\mathbf{x}, q_t = j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \left(\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \Big(x(t) - m_x(j, t) \Big)^2 \right). \tag{14}$$

The re-estimation equations of the rest of the parameters are found by setting the partial derivatives of (14) w.r.t. the parameters to zero, and solving the equation system. Define

[1] In the following, the notation of summation is abbreviated to showing only the variable's name if the summation interval is over the whole range of the variable. In other case the summation interval will be shown explicitly.

$\gamma(j,t) = \frac{p(\mathbf{x}, q_t=j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})}$, which is now interpreted as the posterior of state j at time t given the observation $\mathbf{x}$ and $\tilde{\boldsymbol{\phi}}$. For $g(k)$, we have p equations:

$$\sum_{j}\sum_{t=1}^{T-1}\gamma(j,t)\Big(x(t)-m_x(j,t)\Big)x(t-k)=0, \quad k=1,\cdots,p. \tag{15}$$

For $m_r(j)$, we have M equations:

$$\sum_{t=1}^{T-1}\gamma(j,t)\Big(x(t)-m_x(j,t)\Big)=0, \quad j=1,\cdots,M. \tag{16}$$

For σ^2, we get

$$\widehat{\sigma^2}=\frac{\sum_j\sum_t^{T-1}\gamma(j,t)\Big(x(t)-m_x(j,t)\Big)^2}{\sum_j\sum_{t=1}^{T-1}\gamma(j,t)}. \tag{17}$$

Equation (15) and (16) form $p+M$ coupled linear equations which can be solved analytically, wherein $m_x(j,t)$ is calculated by (5). Then (17) can be solved by inserting the estimated $g(k)$ and $m_r(j)$.

In this model, $m_x(j,t)$ can be interpreted as the linear prediction of $x(t)$ taking into account the mean of the state $q_t=j$. The re-estimation equations also have intuitive interpretations. In (13), a_{ij} equals the expected number of transitions from state i to state j divided by the expected number of transitions made from state i; Equation (15) is a multi-state version of the orthogonality principle; Equation (16) tells that the prediction error weighted by state posterior is of zero mean; and (17) calculates the mean of the prediction error power weighted by the state posterior as the variance of the stochastic element of the signal.

The existence of linear solutions to the maximization of the Q function makes fast convergence. This is a direct benefit from the HMM modeling of the excitation, where the HMM is constrained to have states with identical emission variance. Without this constraint, the resulting maximization equations would be a set of nonlinear equations. GMM-based, general purpose identification methods do not have this constraint, e.g. [11]. Thus they have to resort to numerical maximization of the Q function, which is known as the Generalized EM algorithm.

A GMM with similar constraint can be used in place of the HMM in our signal model, and the EM equations can be derived in the same way as shown above with proper changes in the definition of α and β (the $\xi(i,j,t)$ used in the HMM is not needed in the GMM). The derivation of the GMARM is briefly described in Appendix 5. The advantage of the GMARM is a lighter computational load than that of the HMARM. Whereas, the lack of dynamic modeling makes the GMARM converge slower and es-

timate less accurately than the HMARM when there is a discrete temporal structure in the excitation that is ignored by the GMM, since the GMM is still a static model. Examples of this discrete temporal structure include the impulse train structure in voiced speech signals and Pulse Position Modulation (PPM) signals, and trellis-coded modulation signals. They all have inherent temporal structures that can be well modeled by a state transition matrix. The GMARM on the other hand, can not exploit this useful information in its estimation. For excitations that have no temporal structure, the two algorithms perform similarly.

Remark: An advantage of this two-layer structure is that the AR model extracts the linear temporal structure from the signal, and the HMM takes care of the nonlinear temporal structure overlooked by the AR model. Thus it is a more efficient way of modeling complex temporal structures than using AR model or HMM alone.

2.2 The Extended-HMARM and its identification

In the previous signal model, the output of the AR filter is assumed to be exactly measurable. In many applications, however, measurement noise is inevitable. To be robust against noise, the signal model need to be extended to incorporate a noise model. Assuming stationary white Gaussian measurement noise, we have a new system model whose structure is depicted in Fig. 2. We term this model the Extended-HMARM (E-HMARM).

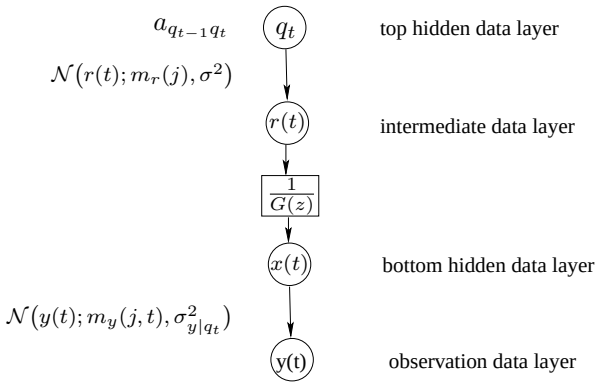


Figure 2: A generative data structure of the E-HMARM.

In this extended data model, we define two hidden data layers: the state q_t and the filter output $x(t)$. Observe that $r(t)$ is not hidden because it is linearly dependent on

$x(t)$. The system model can be expressed in the following equations:

$$x(t) = \sum_{k=1}^{P} g(k)x(t-k) + r(t) \tag{18}$$

$$r(t) = v(t) + u(t) \tag{19}$$

$$y(t) = x(t) + z(t), \tag{20}$$

where $y(t)$ is the observations, $z(t)$ is the measurement noise, $g(k)$ is the kth AR coefficient, and $r(t)$ is the non-Gaussian process noise, or, the filter excitation. We write $r(t)$ as the sum of $v(t)$, a sequence of M-state symbols, and a white Gaussian noise sequence $u(t)$ with zero mean and variance σ_u^2. Thus the excitation $r(t)$ is actually a Hidden Markov process with M states. In HMM terms, these states have Gaussian emission pdfs with mean $m_r(j)$, $j \in [1, \cdots, M]$, and identical variance σ_u^2. The state transition matrix is denoted by $\boldsymbol{A} = (a_{i,j})$. The observation noise is assumed to be white Gaussian noise with zero mean and variance σ_z^2.

The HMM used here is different from the standard HMM and the HMM used in the HMARM in that, the emission pdf of the state $q_t = j$ observed at the observation data layer is a Gaussian pdf with a *time varying mean* $m_y(j,t)$ and a *time varying* variance $\sigma_{y|q_t}^2$. This can be written as:

$$b_{y_t|q_t,\mathbf{y}}(j,t) = \mathcal{N}\Big(y(t); m_y(j,t), \sigma_{y|q_t}^2\Big). \tag{21}$$

From (20), the mean of $y(t)$ should be $x(t)$ if $x(t)$ was known. But since $x(t)$ is not available, a proper choice of the mean of $y(t)$ will be the mean of $x(t)$ given $\mathbf{y}$. So $m_y(j,t)$ can be obtained by calculating the smoothing estimate of $x(t)$ using the observations $\mathbf{y}$ and the current state q_t. The variance of the emission pdf is therefore the sum of the smoothing error variance and the measurement noise variance. The smoothing estimates and the error variance can be calculated with a nonlinear MMSE smoother, which will be described later. It can be summarized as follows:

$$m_y(j,t) = \Big\langle x(t)|\mathbf{y}, \tilde{\boldsymbol{\phi}} \Big\rangle, \tag{22}$$

$$\sigma_{y|q_t}^2 = \sigma_{x_p}^2(j,t) + \sigma_z^2, \tag{23}$$

with $\sigma_{x_p}^2(j,t)$ being the smoothing error variance of $x(t)$ given $q_t = j$. In (22) and in the following, we use the angle bracket $\langle \psi|\varphi \rangle$ to denote the expectation of ψ conditioned on φ. The forward and backward likelihood denoted by $\alpha(j,t)$ and $\beta(j,t)$ are defined in the same way as in the HMARM, and can be calculated recursively.

The parameters to be estimated are $\boldsymbol{\phi} = [\mathbf{A}, m_r(1), \cdots, m_r(M), \sigma_u^2, \sigma_z^2, g(1), \cdots, g(p)]$.

Applying the EM methodology again, we write the Q function as follows:

$$Q(\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}) = \sum_{\mathbf{q}} \int_{\mathbf{x}} p(\mathbf{q}, \mathbf{x}|\mathbf{y}, \tilde{\boldsymbol{\phi}}) \log p(\mathbf{q}, \mathbf{x}, \mathbf{y}|\boldsymbol{\phi}) \mathrm{d}\mathbf{x} \tag{24}$$

$$= \sum_{\mathbf{q}} p(\mathbf{q}|\mathbf{y}, \tilde{\boldsymbol{\phi}}) \log p(\mathbf{q}|\boldsymbol{\phi}) + \sum_{\mathbf{q}} p(\mathbf{q}|\mathbf{y}, \tilde{\boldsymbol{\phi}}) \int_{\mathbf{x}} p(\mathbf{x}|\mathbf{q}, \mathbf{y}, \tilde{\boldsymbol{\phi}}) \log p(\mathbf{x}|\mathbf{q}, \tilde{\boldsymbol{\phi}}) \mathrm{d}\mathbf{x}$$

$$+ \int_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}, \tilde{\boldsymbol{\phi}}) \log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\phi}) \mathrm{d}\mathbf{x}. \tag{25}$$

Equation (25) follows from the first order Markovian property of the layered data model:

$$p(\mathbf{q}, \mathbf{x}, \mathbf{y}|\boldsymbol{\phi}) = p(\mathbf{q}|\boldsymbol{\phi}) p(\mathbf{x}|\mathbf{q}, \boldsymbol{\phi}) p(\mathbf{y}|\mathbf{x}, \boldsymbol{\phi}). \tag{26}$$

Denote the first, second, and third term in (25) as Q_T, Q_B, and Q_V, respectively. Thus Q_T involves only the top hidden layer parameters, Q_B involves only the bottom hidden layer parameters, and Q_V involves only the visible (observation) layer parameters. The maximization of the Q function can now be done by maximizing the three terms in (25) separately.

According to the Gaussian assumption of the observation noise, Q_V can be written as:

$$Q_V = \int_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}, \tilde{\boldsymbol{\phi}}) \sum_t \left[\log \frac{1}{\sqrt{2\pi\sigma_z^2}} - \frac{1}{2\sigma_z^2} \big(y(t) - x(t)\big)^2 \right] \mathrm{d}\mathbf{x} \tag{27}$$

$$= \sum_t \int_{x(t)} p(x(t)|\mathbf{y}, \tilde{\boldsymbol{\phi}}) \left[\log \frac{1}{\sqrt{2\pi\sigma_z^2}} - \frac{1}{2\sigma_z^2} \big(y(t) - x(t)\big)^2 \right] \mathrm{d}x(t) \tag{28}$$

$$= \sum_t \log \frac{1}{\sqrt{2\pi\sigma_z^2}} - \frac{1}{2\sigma_z^2} \sum_t \Big(y^2(t) - 2y(t) \langle x(t)|\mathbf{y} \rangle + \langle x^2(t)|\mathbf{y} \rangle \Big). \tag{29}$$

Note that all the conditioned mean should also be conditioned on $\tilde{\boldsymbol{\phi}}$, but it is omitted here and in the sequel for brevity.

From (18) and (25), Q_B can be written as:

$$\begin{aligned}
Q_B &= \sum_{\mathbf{q}} p(\mathbf{q}|\mathbf{y},\tilde{\boldsymbol{\phi}}) \int_{\mathbf{x}} p(\mathbf{x}|\mathbf{q},\mathbf{y},\tilde{\boldsymbol{\phi}}) \sum_t \Bigg[\log \frac{1}{\sqrt{2\pi\sigma_u^2}} - \frac{1}{2\sigma_u^2}\Bigg(x(t) \\
&\quad - \sum_{k=1}^{P} g(k)x(t-k) - m_r(j)\Bigg)^2\Bigg] \mathrm{d}\mathbf{x} \\
&= \sum_t \sum_j p(q_t|\mathbf{y},\tilde{\boldsymbol{\phi}}) \int_{x(t)} p(x(t)|q_t,\mathbf{y},\tilde{\boldsymbol{\phi}}) \Bigg[\log \frac{1}{\sqrt{2\pi\sigma_u^2}} - \frac{1}{2\sigma_u^2}\Bigg(x(t) \\
&\quad - \sum_{k=1}^{P} g(k)x(t-k) - m_r(j)\Bigg)^2\Bigg] \mathrm{d}x(t) \\
&= \sum_t \sum_j p(q_t|\mathbf{y},\tilde{\boldsymbol{\phi}}) \Bigg[\log \frac{1}{\sqrt{2\pi\sigma_u^2}} - \frac{1}{2\sigma_u^2}\Bigg\langle\Bigg(x(t) \\
&\quad - \sum_{k=1}^{P} g(k)x(t-k) - m_r(j)\Bigg)^2 |q_t,\mathbf{y}\Bigg\rangle\Bigg],
\end{aligned} \tag{30}$$

where $q_t = j$, and $j \in (1,\cdots,M)$. Here, the posterior mean is conditioned on both the state at present time q_t and the observation $\mathbf{y}$.

Q_T can be written as:

$$\begin{aligned}
Q_T &= \sum_{\mathbf{q}} p(\mathbf{q}|\mathbf{y},\tilde{\boldsymbol{\phi}}) \sum_t \log a_{q_{t-1}q_t} \\
&= \sum_t \sum_j p(q_t = j|\mathbf{y},\tilde{\boldsymbol{\phi}}) \log a_{q_{t-1}q_t},
\end{aligned} \tag{31}$$

where $a_{q_{t-1}q_t}$ is the state transition probability (from state q_{t-1} to q_t).

Now we maximize the Q functions by setting the derivatives with respect to the parameters to zeros. For σ_z^2 we get equations:

$$\frac{\partial Q_V}{\partial \sigma_z^2} = -\frac{T}{2\sigma_z^2} + \frac{1}{2(\sigma_z^2)^2} \sum_t \Big[y^2(t) - 2y(t)\big\langle x(t)|\mathbf{y}\big\rangle + \big\langle x^2(t)|\mathbf{y}\big\rangle\Big] \doteq 0,$$

from which we get

$$\sigma_z^2 = \sum_t \Big[y^2(t) - 2y(t)\big\langle x(t)|\mathbf{y}\big\rangle + \big\langle x^2(t)|\mathbf{y}\big\rangle\Big]/T. \tag{32}$$

For the AR parameters $g(k)$, we get:

$$\begin{aligned}\frac{\partial Q_B}{\partial g(k)} &= \sum_{t=1}^{T-1}\sum_{j} p(q_t=j|\mathbf{y},\tilde{\boldsymbol{\phi}})\Big[\frac{1}{\sigma_u^2}\Big\langle\Big(x(t)-\sum_{c=1}^{P} g(k)x(t-c) \\ &\qquad - m_r(j)\Big)x(t-k)\Big|q_t=j,\mathbf{y}\Big\rangle\Big] \\ &= \frac{1}{\sigma_u^2}\sum_{t=1}^{T-1}\sum_{j} p(q_t=j|\mathbf{y},\tilde{\boldsymbol{\phi}})\Big[\Big\langle x(t)x(t-k)|q_t=j,\mathbf{y}\Big\rangle \\ &\qquad - \sum_{c=1}^{P} g(c)\Big\langle x(t-c)x(t-k)|q_t=j,\mathbf{y}\Big\rangle \\ &\qquad - m_r(j)\Big\langle x(t-k)|q_t=j,\mathbf{y}\Big\rangle\Big] \doteq 0, \quad k=1,\cdots,P \end{aligned} \tag{33}$$

Here, $p(q_t=j|\mathbf{y},\tilde{\boldsymbol{\phi}})$ is the posterior probability of the state being j at time t, and is to be denoted in the sequel by $\gamma(t,j)=p(q_t=j|\mathbf{y},\tilde{\boldsymbol{\phi}})$. In (33), the sum of the posterior mean $\langle\ \cdot\ |q_t=j,\mathbf{y}\rangle$ over the state weighted by the state posterior can be expressed as the posterior mean conditioned only on $\mathbf{y}$. That is,

$$\sum_{j}\gamma(t,j)\langle\ \cdot\ |q_t,\mathbf{y}\rangle = \langle\ \cdot\ |\mathbf{y}\rangle. \tag{34}$$

Therefore, (33) can be re-written as

$$\begin{aligned}\sum_{t=1}^{T-1}\Big[&\Big\langle x(t)x(t-k)\Big|\mathbf{y}\Big\rangle - \sum_{c=1}^{P} g(c)\Big\langle x(t-c)x(t-k)\Big|\mathbf{y}\Big\rangle \\ &- \sum_{j}\gamma(t,j)m_r(j)\Big\langle x(t-k)\Big|q_t=j,\mathbf{y}\Big\rangle\Big] = 0.\end{aligned} \tag{35}$$

For $m_r(j)$ we have

$$\begin{aligned}\frac{\partial Q_B}{\partial m_r(j)} &= \sum_{t=1}^{T-1}\gamma(t,j)\Big[-\frac{1}{2\sigma_u^2}\Big\langle 2\Big(x(t)-\sum_{c=1}^{P} g(c)x(t-c)-m_r(j)\Big)(-1)\Big|q_t=j,\mathbf{y}\Big\rangle\Big] \\ &= \frac{1}{\sigma_u^2}\sum_{t=1}^{T-1}\gamma(t,j)\Big[\Big\langle x(t)|q_t=j,\mathbf{y}\Big\rangle - \sum_{c=1}^{P} g(c)\Big\langle x(t-c)|q_t=j,\mathbf{y}\Big\rangle \\ &\qquad - m_r(j)\Big] \doteq 0, \quad j=1,\cdots,M.\end{aligned} \tag{36}$$

For σ_u^2, we have

$$\frac{\partial Q_B}{\partial \sigma_u^2} = \sum_{t=1}^{T-1}\sum_j \gamma(t,j)\Big[-\frac{1}{2\sigma_u^2} + \frac{1}{2(\sigma_u^2)^2}\Big\langle\Big(x(t) - \sum_{c=1}^{P} g(c)x(t-c) - m_r(j)\Big)^2\Big|q_t = j, \mathbf{y}\Big\rangle\Big] \doteq 0, \tag{37}$$

from which we get

$$\sigma_u^2 = \sum_{t=1}^{T-1}\sum_j \gamma(t,j)\underbrace{\Big[\Big\langle\Big(x(t) - \sum_{c=1}^{P} g(c)x(t-c) - m_r(j)\Big)^2\Big|q_t = j, \mathbf{y}\Big\rangle\Big]}_{W} \Big/ \sum_{t=1}^{T-1}\sum_j \gamma(t,j), \tag{38}$$

where

$$\begin{aligned} W = {} & \Big\langle x^2(t)\Big|q_t = j, \mathbf{y}\Big\rangle - 2m_r(j)\Big\langle x(t)\Big|q_t = j, \mathbf{y}\Big\rangle + m_r^2(j) \\ & - 2\sum_{c=1}^{P} g(c)\Big\langle x(t)x(t-c)\Big|q_t = j, \mathbf{y}\Big\rangle + 2m_r(j)\sum_{c=1}^{P} g(c)\Big\langle x(t-c)\Big|q_t = j, \mathbf{y}\Big\rangle \\ & + \sum_{c=1}^{P}\sum_{d=1}^{P} g(c)g(d)\Big\langle x(t-c)x(t-d)\Big|q_t = j, \mathbf{y}\Big\rangle. \end{aligned} \tag{39}$$

The transition probability can be estimated in the same way as in the standard HMM:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} p(\mathbf{x}, q_{t-1} = i, q_t = j|\tilde{\boldsymbol{\phi}})}{\sum_{t=1}^{T-1} p(\mathbf{x}, q_{t-1} = i|\tilde{\boldsymbol{\phi}})} = \frac{\sum_{t=1}^{T-1}\xi(i,j,t)}{\sum_{t=1}^{T-1}\gamma(i,t)}, \tag{40}$$

where $\xi(i,j,t)$ and $\gamma(i,t)$ are defined in the same way as in the HMARM.

Equation (32), (35), and (36) consist of a set of $1 + P + M$ linear equations and can be solved by matrix inversion. Then (38) can be solved by inserting the newly updated parameter estimates. The quantities needed in these equations include: the state posteriors $\xi(i,j,t)$ and $\gamma(i,t)$, which are calculated by the forward-backward algorithm; the first and second moments of $x(t)$, which are estimated by a nonlinear MMSE fixed-interval smoother.

The nonlinear MMSE smoother consists of a forward sweep and a backward sweep. In the forward sweep, at time t, a Kalman filter produces M estimates of the mean and correlation matrix of $x(t)$ conditioned on $q_t = j$, $j = 1, \cdots, M$, and $\mathbf{y}$. We com-

bine the M estimates weighted by the state *a posteriori* probabilities, $\gamma(i,t)$, to get an MMSE filtering estimate conditioned only on $\mathbf{y}$. Then the backward sweep calculates the smoothing estimates and MSE matrices using the filtering estimates and MSE matrices obtained in the forward sweep. The backward sweep equations are identical to those of the two-pass Kalman smoother, and can be found in, e.g., [12, p.572]. The algorithm thus iterates between the nonlinear MMSE smoother, and the estimation of $\boldsymbol{\phi}$ and $\gamma(i,t)$.

The algorithm stacks two dynamic state estimators together, i.e., the nonlinear MMSE smoother and the HMM estimator. A unifying view of the Kalman-type state estimator and the HMM state estimator can be found in [13]. The nonlinear smoother uses a continuous state model, where the state vector is the output of the AR(P) filter, $\mathbf{x}|_{t-P+1:t}$, and the state transition is ruled by the auto-regressive property of the AR(P) filter. The HMM uses a discrete state model, where the states are the input symbols, and the state transition is ruled by the underlying mechanism that produces the symbols.

Remark: The proposed nonlinear MMSE smoother falls in the category of Switching Kalman Filter (SKF) with soft-decision, as is defined in [4]. Different from the typical SKFs whose control mechanism switches the AR filter coefficients and/or the system-noise variance over segments of data, the proposed SKF switches its system-noise mean from sample to sample.

3 Applications and results

We apply the proposed system models and their identification algorithms to tackle problems in speech analysis and channel equalization. In the speech analysis examples, we show that the proposed non-Gaussian AR system identification method can provide better estimates of the AR coefficients, and better structured residual, than those given by the classical LPC analysis. We also show that under mild noise conditions, robust AR analysis can be achieved without knowing the noise variance. In the channel equalization examples, we show that joint channel estimation and symbol estimation can be done efficiently to a high accuracy when SNR is high. When SNR is moderate, the joint estimation can be done with extra computational complexity.

3.1 Efficient non-Gaussian speech analysis

In a vast variety of speech processing applications, AR coefficients or AR spectra, and linear prediction residual need to be calculated. Least Squares methods, such as the LPC analysis (implemented as an autocorrelation method), have been the standard methods of analyzing AR models. The Gaussian assumption taken by the LS method results in simple analytic solutions. But when applied to non-Gaussian signals such as voiced speech signals, the mismatch of assumption brings in undesirably large variance and

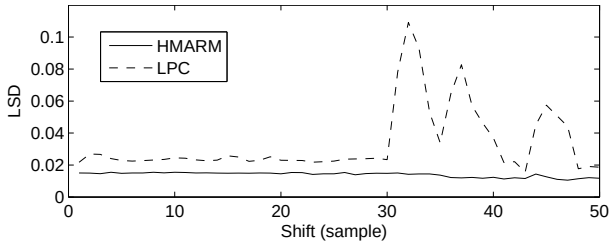
bias. The large variance implies a bad shift-invariance property of the LPC analysis. This means that, when a sustained vowel is segmented into several frames, the LPC estimates of the AR parameters for each frame can be very different. This causes, as an example, in a CELP coding application, more bits than necessary to be transmitted, and in a packet loss concealment application, difficulty to interpolate a missing frame. Here we apply the HMARM method to AR analysis, and compare the bias and the variance of the estimates to those given by the LPC analysis.

First, we use a synthetic signal that resembles a sustained voiced speech signal. The synthetic signal is made by filtering a noisy impulse train with an AR(10) filter. 50 realizations of this signal are analyzed. To get the 50 realizations we shift a rectangular window along the signal one sample each time 50 times. The window length is 320 samples. The estimated AR spectra of the 50 realizations are compared to the true AR spectrum, and the difference is measured by the Log-Spectral Distortion (LSD) measure. The LSD is defined as follows:

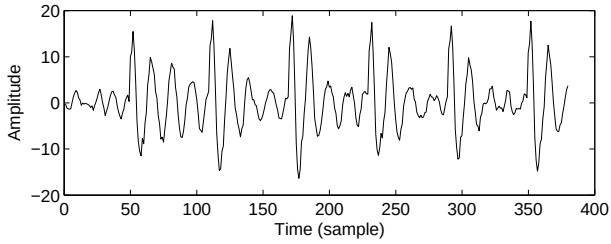
$$LSD = \frac{1}{L}\left[\sum_{l=1}^{L}\left(20log_{10}\frac{|X(l)|}{|\hat{X}(l)|}\right)^2\right]^{\frac{1}{2}}, \tag{41}$$

where L is the number of spectral bins. The LSD versus the shift is shown in Fig 3. It is clear that the proposed method has a flat distortion surface and this surface is lower than the LPC's. It is important to note that the LPC estimates encounter huge deviation from the true values in the second half of the plot. This is where a large "hump" in the signal comes into the analysis frame. The large humps in the signal are caused by the impulses in the excitation, which represent the non-Gaussian/nonlinear structure of the signal. The bias and variance of the estimates are also calculated using sample mean and sample variance. The bias is 0.092 for the HMARM analysis, and compared to the 0.197 for the LPC analysis, accounts for an improvement of more than 5 dB. The variance is 0.128 for the HMARM and 9.69 for the LPC analysis, representing a variance reduction of 18.8 dB.

Now, we test the shift-invariance property with true speech signals. For real speech signals, there is an implementation issue needed to be pointed out. Since the HMARM is a causal dynamic model, and the analysis is usually frame-based, the ringing from the last impulse of the previous frame has an undesired impact on the current frame estimates. This is because the estimator does not see the previous impulse but its effect exists. This could sometimes degrade the performance mildly. We therefore suggest to do a pre-processing that removes the ringing from the previous frame, or simply set the signal before the first impulse to zero. The latter is used in our experiments. The AR spectra of four different voiced phonemes are estimated 50 times with one sample shift each time. The frame length is set to 256 samples. The spectra are plotted in Fig 4. The estimates by the HMARM show good consistency, while the consistency of the



(a)



(b)

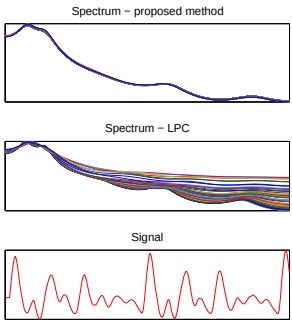
Figure 3: (a): The Log-Spectral Distortion of the AR spectra of the 50 shifted frames. (b): the synthetic signal waveform used in the experiment.

LPC analysis appears to be poor. We observed the same tendency when we varied the segment length and compared the estimates from different data length. These results show that, the LPC analysis is sensitive to the difference in the waveforms of different realizations of the same process, while the HMARM is significantly less sensitive. The residual of the HMARM analysis also has different properties than the LPC analysis. In Fig. 5 we show the prediction residual of a voiced speech signal using the AR parameters estimated by the HMARM and the LPC respectively. It is clear that the residual of the HMARM has more prominent impulses, and the noise between the impulses appears to be less correlated. In general, the residual of HMARM has a smaller L1 norm than that of the LPC analysis. From a sparse coding point of view, the proposed method provides a sparser representation of the voice signal than the one given by LPC analysis. Traditionally, sparse representation is achieved by minimizing L1-norm with numerical optimizations (see [14] for a review, and [15] for application in speech analysis), or using Bayesian inference with a super Gaussian pdf as prior [16]. The HMARM method proposed here provides a computationally simple alternative to the sparse coding of voiced speech signals.

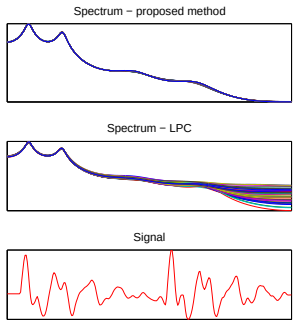
In the experiments described above, the analysis window is a rectangular window. As it is well known that an appropriately chosen window can reduce the variance of

the LPC estimates, we also conducted comparisons between the HMARM analysis and the Hamming-windowed LPC analysis. For the synthetic signal, the variance of the Hamming-windowed LPC is 1.197, which is still 9.7 dB higher than that of the HMARM. Although its variance is reduced, the Hamming-windowed LPC in general suffers from lower spectral resolution due to the large main lobe of the Hamming window. We show in Fig. 6 that the Hamming-windowed LPC analysis fails to resolve two closely located spectral peaks, while the HMARM succeeds. The signal used herein is a synthetic signal, which is made by filtering a noisy impulse train with an AR filter with order 40. Windowing technique can sometimes cause large bias because it alters the signal waveform significantly, especially when the data sequence is short. We show in Fig. 7 the difference in spectrum caused by windowing. By reducing the amplitude of the last peak, the Hamming window changes the waveform and thus the spectrum significantly.

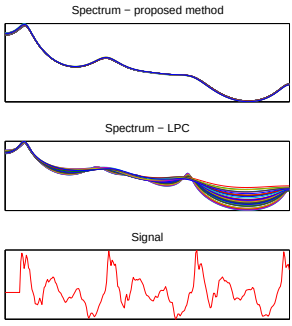
Another known LS method is the covariance method [17, Ch. 5.3]. The covariance method is known to give more accurate estimates of the AR coefficients than the autocorrelation method when the data length is small. In our experiments, it is so when the analysis window is rectangular. When a Hamming window is used, the covariance method gives similar results as the autocorrelation method.



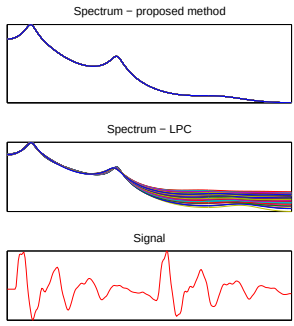
(a) Vowel /o/ in “short”.



(b) Vowel /e/ in “red”.



(c) Vowel /a/ in “began”.



(d) Vowel /ou/ in “shoulder”

Figure 4: The AR spectra estimated by HMARM and LPC analysis.

3.2 Blind channel equalization

We consider a discrete-time communication channel model as shown in Fig. 8, where the channel response has included the response of the transmitter filter, the medium, the receiver filter, and the symbol-rate sampler. We assume that the channel can be well characterized by an AR model, and no measurement noise is present (or, the channel has a very high SNR). The transmitted symbols are quaternary PAM symbols. At the receiver end, the channel distortion is compensated and the transmitted symbols are decoded. The receiver has no prior knowledge about the channel, the alphabet of the transmitted symbols, and the probability distribution of the symbols.

Using the HMARM, the equalization and decoding are done jointly. In the first experiment, 200 symbols generated randomly using a four-symbol alphabet $\mathcal{A}$={ -3,-

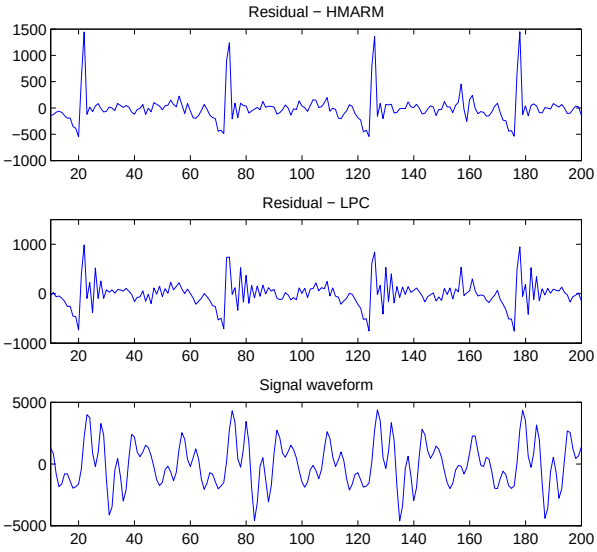


Figure 5: Prediction residuals by the HMARM and the LPC analysis.

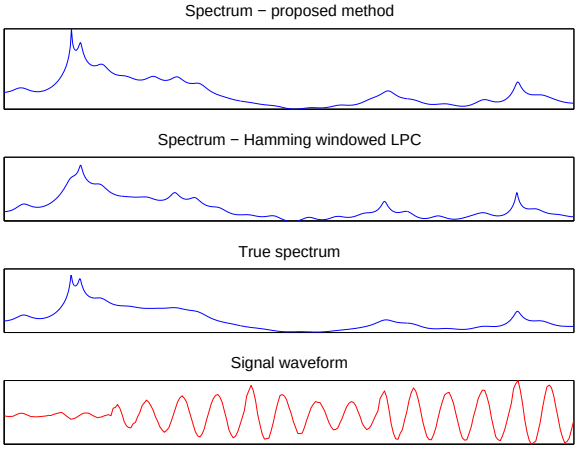


Figure 6: Spectral resolution comparison using a synthetic signal. The AR model order is 40.

1,1,3} are transmitted. The channel is AR(10) with coefficients

$$A = [1, -1.223, -0.120, 1.016, 0.031, -0.542, -0.229, 0.659, 0.307, -0.756, 0.387].$$

The received signal waveform, the equalizer output, and the estimated channel spectra

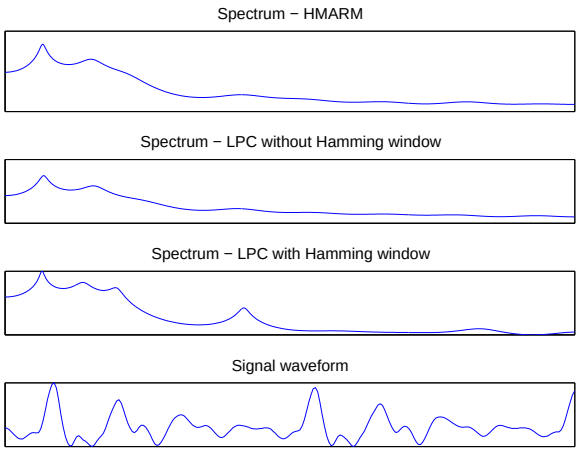


Figure 7: Using Hamming window on a short frame alters the spectrum.

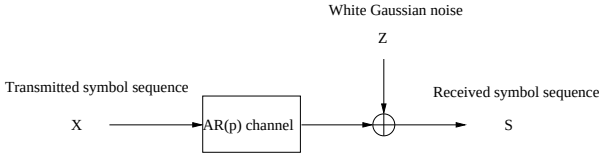


Figure 8: The discrete-time channel model.

are shown in Fig. 9 Fig. 10 and Fig. 11, respectively. Here we again use the LS method as the reference method. It is clear from the figures that the recovered symbol sequence by the HMARM method coincides with the transmitted symbols very well, and the spectrum estimated by the HMARM method completely overlaps with the true channel spectrum. Whereas the LS method has a much larger estimation error on both the recovered symbols and the channel spectrum. More precisely, the estimation error variance of the recovered symbol sequence is 1.06×10^{-26} for the HMARM method and 0.36 for the LS method, which represents a 255 dB gain of the HMARM method over the LS method.

In the second experiment, we consider an FIR channel model. In most of the channel equalization literature, channels are modeled by MA models. A major advantage of MA modeling in channel equalization is the simplicity in algorithm design. Whereas, most realistic channels have both an MA part and an AR part. When the channel response is IIR, the drawback of an MA model is obvious: it requires a very large number of coefficients to approximate an IIR channel, while the AR model can approximate an MA channel with a mildly larger order. Equalization of MA channel using AR model

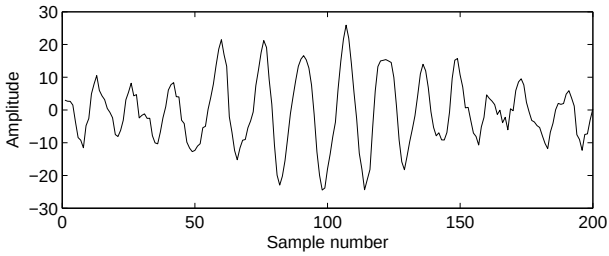


Figure 9: The received signal waveform. The channel is AR(10).

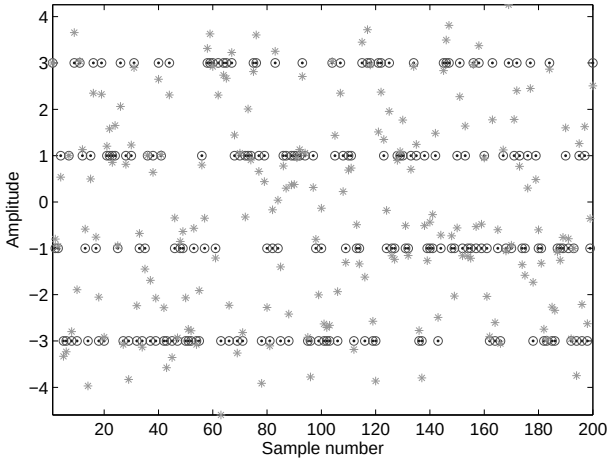


Figure 10: The recovered symbol sequences. Dots: the transmitted symbols, circles: the recovered symbols by the HMARM, stars: the recovered symbols by the LS method.

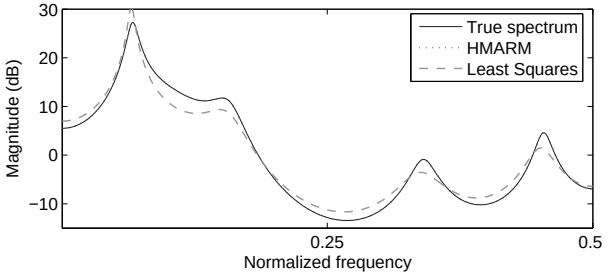


Figure 11: The true and estimated spectra. Note that the HMARM spectrum overlaps the true spectrum.

has been shown before, e.g. [11]. In this example we use the same experimental setup as in [11] to demonstrate the applicability of our method to the MA channel equalization. The alphabet $\mathcal{A}$ is the same as before, and the 3rd order MA channel coefficients

are $B = [1.0, -0.65, 0.06, 0.41]$. The received signal waveform is shown in Fig. 12. The recovered symbol sequence and the estimated channel spectrum are shown in Fig. 13 and Fig. 14, respectively. The estimation error variance of the recovered symbol sequence is 0.0023 for the HMARM method, and 0.4212 for the LS method. The gain of the HMARM method over the LS method is 22.6 dB.

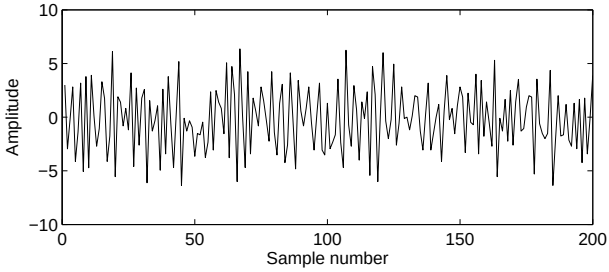


Figure 12: The received signal waveform. The channel is MA(3).

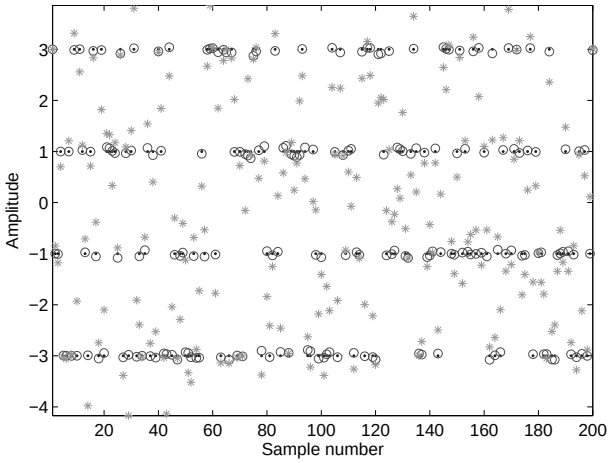


Figure 13: The recovered symbol sequences. Dots: the transmitted symbols, circles: the recovered symbols by the HMARM, stars: the recovered symbols by the LS method.

When there exists white Gaussian measurement noise in the system, the performance of the HMARM method degrades. For a channel SNR of 60 dB, 50 dB, and 40 dB, the gain of the HMARM method over the LS method are 27.5 dB, 17.5 dB, and 8 dB, respectively. From 30 dB down, the performance of HMARM is similar to that of the LS method.

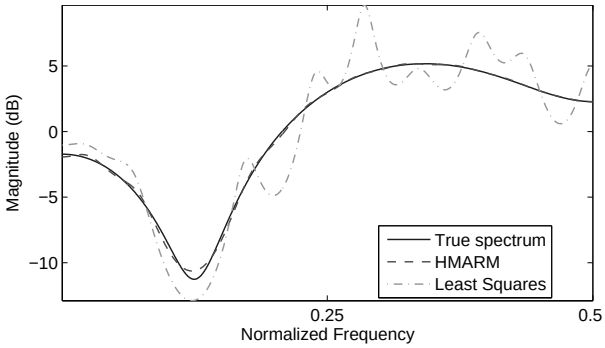


Figure 14: The true and estimated spectra.

3.3 Noise robust spectrum estimation for voiced speech

When measurement noise is present in an AR system, the classic Least Squares method performs poorly because there is no noise modeling in it. The LS method can be extended to modeling both process noise and measurement noise. This is known as the Extended Least Squares (XLS) method [18]. Examples of Gaussian AR model identification are given in [18]. On another thread, EM-type AR model estimations in noisy environments have been extensively studied, especially in the speech processing literature. Pioneered by Lim and Oppenheim [5], and followed by Hansen and Clements [19], Weinstein and Oppenheim [20], Gannot [21], and etc., the paradigm of EM-type algorithms is an iterative ML or MAP estimation. These algorithms are all based on Gaussian signal assumption and succeed in achieving noise robust AR estimation with low complexities. Yet a common drawback of the Gaussian EM-type algorithm is that convergence is not guaranteed. Often an empirical stop criterion is needed, or certain constraints based on knowledge of speech signals are needed [19].

Using the E-HMARM method, we show that the observation noise strength, the AR parameters, and the excitation statistics of voiced speech signal can be jointly estimated, and the convergence is guaranteed.

The synthetic signal used in Section 3.1 is added with white Gaussian noise, such that the SNR equals 15 dB and 20 dB. Fig. 15 and Fig. 16 show the signal spectrum and the estimated spectra by the E-HMARM and LS, respectively. Table 1 shows the averaged values of parameters of 50 estimations. The results show that the E-HMARM algorithm gives much better estimates of the signal spectra than the LS method. The estimates of the impulse amplitude and measurement noise variance are also quite accurate. The estimated process noise variance is always larger than the true value, especially when the SNR is low. This is because in the E-HMARM algorithm, the modeling error is included as part of the process noise.

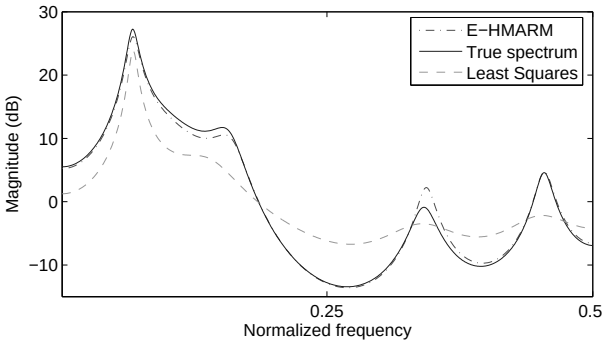


Figure 15: The true and estimated spectra. The SNR is 15 dB.

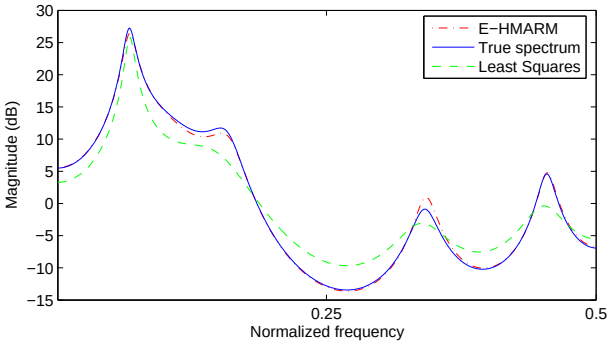


Figure 16: The true and estimated spectra. The SNR is 20 dB.

Table 1: The true and estimated parameters. Results are the average of 50 estimations.

	AR(10) filter coefficients	σ_z^2	σ_u^2	$m_r(1)$	$m_r(2)$
True values	[1, -1.223, -0.120, 1.016, 0.031, -0.543, -0.229, 0.660 , 0.307, -0.756, 0.387]	1.43[a] 0.45[b]	0.22	0	10
E-HMARM (15 dB)	[1, -1.164, -0.144, 0.916, 0.078, -0.438, -0.286, 0.601, 0.316, -0.678, 0.334]	1.36	0.57	-0.03	10.25
E-HMARM (20 dB)	[1, -1.210, -0.126, 0.975, 0.044, -0.475, -0.290, 0.646, 0.335, -0.759, 0.375]	0.51	0.27	-0.03	10.23
LS (15 dB)	[1, -0.838, -0.150, 0.425, 0.184, -0.041, -0.119, 0.109, 0.237, -0.098, 0.075]	-	-	-	-
LS (20 dB)	[1, -1.012, -0.143, 0.650, 0.140, -0.211, -0.200, 0.266, 0.327, -0.341, 0.155]	-	-	-	-

[a] 15 dB
[b] 20 dB

Like all EM-type algorithms, it is possible for the E-HMARM algorithm to con-

verge towards a local maxima. A good initialization can prevent converging to the local maxima. In our implementation of the E-HMARM algorithm, the LS estimates of the AR coefficients are used as initial values. The convergence criterion is set such that the iteration stops when the norm of the difference in the parameter vectors is smaller than 10^{-4}. No divergence has ever been observed under extensive experiments. The E-HMARM algorithm works best at SNRs above 15 dB. From 10 dB and below, the algorithm converges to the LS solution.

3.4 Blind noisy channel equalization

In Section 3.2 we have shown the performance of the HMARM blind channel equalization in a high SNR communication system. We now show that at a lower SNR range, the E-HMARM algorithm can do the job better.

In this example, we consider a Pulse Position Modulation (PPM) signal. PPM is a modulation scheme in which M messages are encoded by transmitting a single pulse in one of M possible time-shifts in a time frame. PPM is typically used in optical communications and recently in ultra-wide-band (UWB) systems [22] and indoor infrared communications [23]. PPM is known to be vulnerable to ISI because of its very large signal bandwidth, and equalization is necessary for high speed transmission. Different from the white spectrum of the PAM symbol sequence, the spectrum of a PPM symbol sequence is high pass and has a strong DC component [2]. The smaller the M, the more high pass the spectrum. This imposes a difficulty to the system identification, i.e., the auto-correlation in the symbol sequence can be absorbed into the AR spectrum estimates resulting in biased estimates of the channel response. In the E-HMARM, this difficulty can be circumvented by exploiting the known symbol amplitudes. That is, if the transmitted symbol amplitudes are known to the receiver, as is the case in most communication systems, we can constrain the m_r to be equal to the known values. This not only speeds up the convergence, but also makes the algorithm robust against the non-whiteness of the symbol sequence.

In the experiment, the transmitted symbols are randomly generated from an M-ary alphabet with $M = 8$. A signal frame thus has 8 time slots, each corresponding to one symbol in the alphabet. When the kth symbol is to be transmitted, a pulse is put at the kth time slot, and zeros elsewhere. We again use an equivalent discrete-time channel model to simplify the analysis. Without loss of generality, the transmitted signal is modeled as a "1" at the symbol position and "0" at the other 7 positions. The channel is modeled as an AR(10) filter. White Gaussian noise is added to the output of the AR(10) filter. The E-HMARM equalizer estimates the channel response and the noise variance,

[2] Instead of defining the whole frame as a symbol, here we treat the pulse duration as the symbol duration. Thus a time frame consists of M symbols, and the sampler at the receiver samples M times per frame. This is why the received symbol sequence has a strong DC component and a high pass spectrum.

and does inverse filtering to recover the transmitted symbols. The standard LS method is used as a reference method. It is shown in Fig. 17 that the recovered symbol sequence by the E-HMARM method has much smaller error variance than that of the LS method. In Fig. 18 it is shown that the E-HMARM gives a very good estimate of the channel spectrum, while the LS estimate is far off. The channel SNR in this example is 18 dB, and the signal length is 400 samples. The E-HMARM equalizer works best at SNRs above 18 dB. At SNRs below 18 dB its performance degrades fast. At SNRs below 15 dB the E-HMARM algorithm converges to the LS solution.

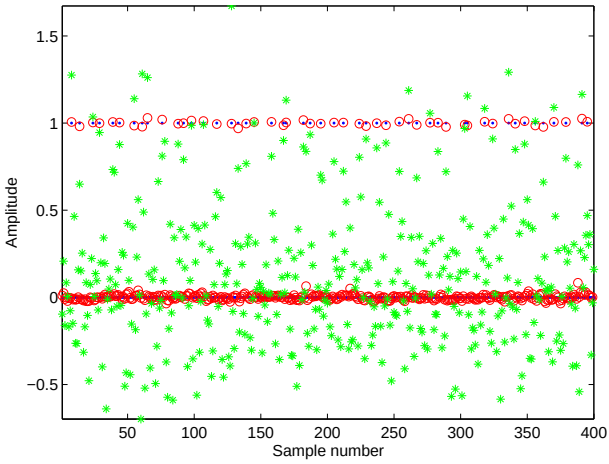


Figure 17: The recovered symbol sequence. Dots: the transmitted symbols, circles: the recovered symbols by the HMARM, stars: the recovered symbols by the LS method. The SNR is 18 dB.

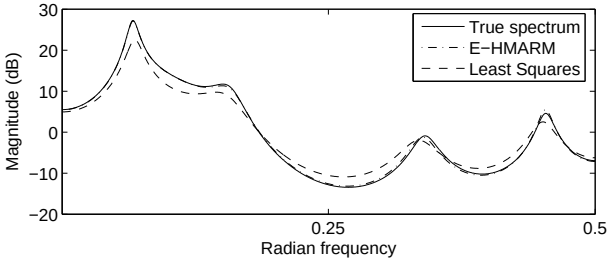


Figure 18: The true and estimated spectra. The SNR is 18 dB.

Next, we consider a combined PAM-PPM modulation with a smaller M. A time frame has now $M = 4$ pulse positions. Only one of the positions has an impulse, and the other positions have zeros. The impulse can have an amplitude of either "1" or "2". So the alphabet still has 8 symbols, but the time frame is shorter and thus the high

pass effect of the symbol sequence is more severe. Fig. 19 shows the spectrum of a transmitted symbol sequence. The LS equalizer mistakes the high pass characteristics of the transmitted symbol sequence as part the channel distortion, and results in a biased spectrum estimate, as shown in Fig. 20. In the same figure, the spectrum estimate by the E-HMARM method is shown, and its curve overlaps the true spectrum. Fig. 21 shows the recovered symbol sequence. It shows clearly that the E-HMARM gives a much lower estimation error variance than the LS method. In this experimental setup, the E-HMARM works best at SNRs above 23 dB.

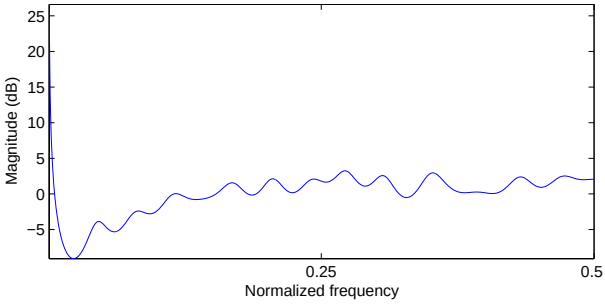


Figure 19: The spectrum of the transmitted symbol sequence.

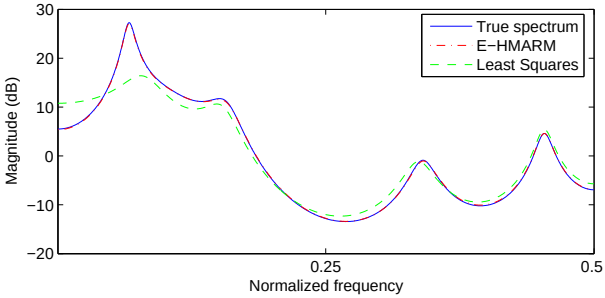


Figure 20: The true and estimated spectra. The SNR is 23 dB.

4 Conclusion

In this paper we have presented two blind system identification algorithms for two non-Gaussian AR systems. The algorithms combine an AR model and an HMM such that second order temporal structure (auto-correlation) and higher order temporal structure (abrupt changes and discrete dynamics) in the signals can be extracted efficiently by the

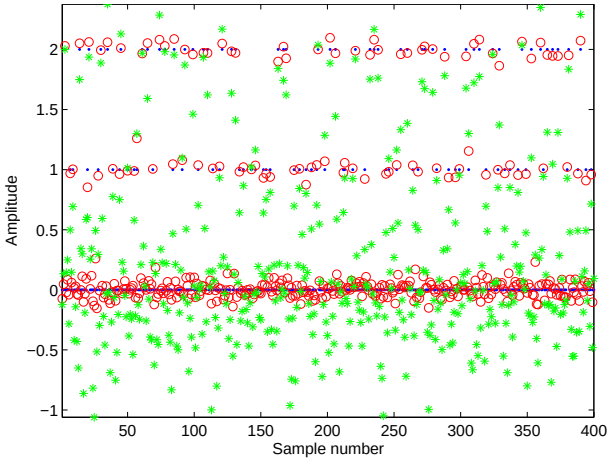


Figure 21: The recovered symbol sequence. Dots: the transmitted symbols, circles: the recovered symbols by the HMARM, stars: the recovered symbols by the LS method. The SNR is 23 dB.

two models, respectively. By constraining the variance of the HMM emission pdfs to be identical, the algorithms have analytical solutions to the maximization of the Q functions in each iteration, which results in faster convergence than numerical optimization methods. In the case that measurement noise is present, a nonlinear MMSE smoother is integrated into the EM algorithm. This smoother obtains optimum MMSE estimates of the non-Gaussian signal at a complexity comparable to M Kalman smoothers. At moderate noise levels, the algorithm gives accurate estimates of the parameters of the HMM, the AR model, and the measurement noise variance. Applications of the algorithms in speech analysis and channel equalization are demonstrated.

5 Appendix I

Here we show how to combine a GMM with an AR model in the two-layer data structure. The forward-backward algorithm used in the HMM parameter learning is a convenient and insightful way of calculating the state posterior probability. So we can modify the HMM learning algorithm to obtain a GMM learning algorithm.

Assume the GMM has M Gaussian terms. Denote the vector of the weights for Gaussian terms by $A = [a_i]$, where $i \in 1, \cdots, M$. Denote the emission pdf given the state $q_t = j$ by $b_x(j,t)$. Define the forward and backward likelihood $\alpha(j,t)$ and $\beta(i,t)$ as same as in the HMM. So the induction equations can be written, analogous to those

of the HMM, as:

$$\begin{aligned}\alpha(j,t) &= \sum_i \Big[\alpha(i,t-1)\Big] a_j b_x(j,t) \\ &= a_j b_x(j,t),\end{aligned} \tag{42}$$

and

$$\beta(i,t) = \sum_j \Big[a_i b_x(j,t+1)\beta(j,t+1)\Big]. \tag{43}$$

Now, we can derive the EM algorithm. The Q function can be written as

$$Q(\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}) = \sum_{\mathbf{q}} p(\mathbf{q}|\mathbf{x}, \tilde{\boldsymbol{\phi}}) \log p(\mathbf{x}, \mathbf{q}|\boldsymbol{\phi}) \tag{44}$$

$$= \sum_{\mathbf{q}} \frac{p(\mathbf{x}, \mathbf{q}|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \left(\sum_{t=1}^{T} \log a_{q_t} + \sum_{t=1}^{T} \log b_x\big(q_t, x(t)\big) \right) \tag{45}$$

$$= \sum_j \sum_{t=1}^{T} \frac{p(\mathbf{x}, q_t = j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \log a_{q_t} + \sum_j \sum_{t=1}^{T} \frac{p(\mathbf{x}, q_t = j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \log b_x\big(q_t, x(t)\big). \tag{46}$$

Comparing (46) with (12), only the first terms are different. So all the re-estimation equations are identical except for the one for a_j. For a_j we have the following re-estimation equation:

$$\begin{aligned}\hat{a}_j &= \frac{\sum_{t=1}^{T} p(\mathbf{x}, q_t = j|\tilde{\boldsymbol{\phi}})}{\sum_j \sum_{t=1}^{T} p(\mathbf{x}, q_t = j|\tilde{\boldsymbol{\phi}})} \\ &= \frac{\sum_{t=1}^{T} \alpha(j,t)\beta(j,t)}{\sum_j \sum_{t=1}^{T} \alpha(j,t)\beta(j,t)}.\end{aligned} \tag{47}$$

This GMARM algorithm has a lighter computational load than the HMARM presented in Section 2.1 since the calculation of the state posterior probability has a simpler form.

References

[1] C. Li and S. V. Andersen, "Blind identification of non-Gaussian Autoregressive models for efficient analysis of speech signals," *Proceedings of ICASSP*, 2006.

[2] D. Sengupta and S. Kay, "Efficient estimation of parameters for non-Gaussian autoregressive processes," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37. No.6, pp. 785–794, 1989.

[3] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[4] K. Murphy, "Switching Kalman filters," *Technical report, U. C. Berkeley*, 1998.

[5] J. S. Lim and A. V. Oppenheim, "All-pole Modeling of Degraded Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASP-26, pp. 197–209, June 1978.

[6] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement," *IEEE Trans. on Signal Processing*, vol. 39, pp. 1732–1742, 1991.

[7] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-Interscience-IEEE, 1993.

[8] L. R. Rabiner and B. H. Juang, "An introduction to Hidden Markov Model," *IEEE ASSP Magazine*, pp. 4–16, Jan. 1986.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc., Series B*, p. 138, 1977.

[10] L. E. Baum and T. Petrie, "Statistic inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.*, vol. 37, pp. 1554–1563, Mar. 1966.

[11] S. M. Verbout, J. M. Ooi, J. T. Ludwig, and A. V. Oppenheim, "Parameter estimation for autoregressive Gaussian-Mixture processes: the EMAX algorithm," *IEEE Trans. on Signal Processing*, vol. 46. No.10, pp. 2744–2756, 1998.

[12] G. Strang and K. Borre, *Linear Algebra, Geodesy and GPS*. Wellesley-Cambridge, U.S., 1997.

[13] S. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Neural Computation*, vol. 11. No.2, 1999.

[14] Y. Li, A. Cichocki, and S.-I. Amari, "Analysis of sparse representation and blind source separation," *Neural Computation*, vol. 16, pp. 1193–1234, 2004.

[15] M. Namba, H. Kamata, and Y. Ishida, "Neural Networks Learning with L1 Criteria and Its Efficiency in Linear Prediction of Speech Signals," *Proc. ICSLP '96*, vol. 2, pp. 1245–1248, 1996.

[16] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1," *Vision Research*, vol. 37, No.23, pp. 3311–3325, 1997.

[17] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*. Wiley Interscience, 2002.

[18] A. Yeredor, "The extended least squares criterion: minimization algorithms and applications," *IEEE Trans. on Signal Processing*, vol. 49. No.1, pp. 74–86, 2000.

[19] J. H. L. Hansen and M. A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, 1991.

[20] E. Weinstein, A. V. Oppenheim, and M. Feder, “Signal enhancement using single and multi-sensor measurements,” *RLE Tech. Rep. 560, MIT, Cambridge, MA*, vol. 46, pp. 1–14, 1990.

[21] S. Gannot, D. Burshtein, and E. Weinstein, “Iterative and sequential Kalman filter-based speech enhancement algorithms,” *IEEE Trans. on Speech and Audio*, vol. 6, pp. 373–385, July 1998.

[22] K. Siwiak, P. Withington, and S. Phelan, “Ultra-wide band radio: The emergence of an important new technology,” *IEEE Proc. Veh. Tech. Conf.*, vol. 2, pp. 1169–1172, 2001.

[23] M. D. Audeh, J. M. Kahn, and J. R. Barry, “Performance of Pulse-Position Modulation on measured non-directed indoor infrared channels,” *IEEE Trans. Communications*, vol. 44, No. 6, pp. 654–659, 1996.

Paper G

Efficient Implementation of the HMARM Model Identification and Its Application in Spectral Analysis

Chunjian Li and Søren Vang Andersen

The paper has been submitted to
Proceedings, 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing.
2007.

The layout has been revised.

Abstract

The Hidden Markov Auto-Regressive model (HMARM) has recently been proposed to model non-Gaussian Auto-Regressive signals with hidden Markov-type driving noise. This model has been shown to be suitable to many signals, including voiced speech and digitally modulated signals received through ISI channels. The HMARM facilitates a blind system identification algorithm that has a good computational efficiency and data efficiency. In this paper, we solve an implementation issue of the HMARM identification, which can otherwise degrade the efficiency of the model and hinder extensive evaluations of the algorithm. Then we study in more detail the properties associated with the autoregressive (AR) spectral analysis for signals of interest.

1 Introduction

Exploiting the non-Gaussianity of signals in spectral analysis can often offer significant improvements in estimation accuracy over traditional Gaussianity based methods. In [1] and [2], we show that specially designed non-Gaussian models for specific types of signals can exploit the structures in the signals and achieve higher computational and data efficiency than general purpose non-Gaussian methods such as the higher order statistics methods and Gaussian Mixture Model based methods. The Hidden Markov Auto-Regressive model (HMARM) proposed by the authors in [1] is tailored for signals generated by exciting an autoregressive (AR) filter with either a finite-alphabet symbol sequence or a hidden Markov sequence. Due to the non-Gaussian nature of the excitation, this type of signal belongs to the class of non-Gaussian AR signals. We proposed an efficient learning algorithm for the HMARM to jointly estimate the AR coefficients and the excitation symbols or the parameters of the hidden Markov sequence. The joint estimation is what distinguishes the method from other identification algorithms of models that have similar source-filter structure: most known methods estimate the source parameters and the filter parameters in a sequential way, resulting in lower efficiencies. The HMARM algorithm is an exact EM algorithm, which solves for a set of linear equations iteratively and converges in a few iterations. It is shown that compared to the classical autocorrelation method of AR spectral analysis, the HMARM has a smaller bias, a smaller variance, and a better shift invariance property. In [2], the HMARM is extended for robust analysis of noisy signals by introducing an observation noise model to the system. At moderate noise levels, the algorithm achieves a high estimation accuracy without *a priori* knowledge of the noise variance. Applications of the model to different signals, including noise robust spectral analysis of speech signals and blind channel estimation, are demonstrated in [1] [2], and promising results are obtained.

One critical issue in the frame based implementation of the HMARM algorithm

in [1] is that, if a signal is segmented into frames, the HMARM could have problems estimating the parameters for those frames that do not contain the onset of the signal. This is because when estimating the AR parameters of the current frame, the estimator has no knowledge about the excitation in the previous frame, but the large impulses in the previous excitation can cause large "ripples" in the beginning of the current frame, which then causes the state estimator in the HMARM to make wrong decisions. Since the parameter estimations are based on the state decisions, these estimates become erroneous too. In the previous papers, this problem is solved by pre-processing the frame to remove the "ripples" caused by the previous frame. For simplicity of that approach, all samples before the first impulse in the current frame are set to zero. This solution is somewhat troublesome since it requires an impulse detector in the residual domain, whose accuracy affects the performance of the whole system. This and other ways of subtracting the ripples also lower the computation efficiency and data efficiency, since they add extra complexity and discard data samples. In this paper, we address this problem by exploiting the Markovian property of the AR model in a way analogous to the covariance method for AR spectral analysis. Our proposed solution costs no extra complexity, and is highly reliable.

The rest of the paper is organized as follows. Section 2 describes the covariance implementation, and discusses its benefits. Then, in Section 3, we investigate some interesting properties of the HMARM using our new proposed implementation in application to spectral analysis.

2 Covariance method for the HMARM

The causality problem associated with the frame based implementation[1] of the HMARM is functionally different from the boundary problem in the least-squares (LS) method. The classical LS solution to the AR spectral analysis assumes the excitation to the AR filter to be a stationary white Gaussian sequence. With this assumption, the only parameter of the excitation statistics, the variance, is decoupled from the estimation of the AR filter coefficients. Therefore, the excitation has no effect on the AR filter estimates. However, the HMARM has a more sophisticated model for the excitation, and the estimations of the excitation parameters and the AR parameters affect each other. Specifically, the HMARM models the excitation as a hidden Markov sequence. During the estimation, the states of the excitation sequence at each time instant are first estimated by calculating the state probabilities. Based on the state decisions, the AR filter coefficients and the parameters of the hidden Markov model are estimated by a set of coupled linear equations, c.f. [1] and [2] for derivations. For convenience, we list below the signal model and the final equations of the estimator.

[1] In this context, the frames have no overlap.

For a signal generated by the following model,

$$x(t) = \sum_{k=1}^{p} g(k)x(t-k) + r(t) \tag{1}$$

$$r(t) = v(t) + u(t), \tag{2}$$

where $x(t)$ is the signal, $g(k)$ is the kth AR coefficient, and $r(t)$ is the excitation sequence consisting of a Markovian sequence $v(t)$ and additive white Gaussian noise $u(t)$, the estimates of the parameters are obtained from solving the following $p + m$ equations, where p is the order of the AR model, and m is the number of states of the HMM. For $k = 1, \cdots, p$, and $j = 1, \cdots, m$:

$$\sum_{j}^{m} \sum_{t=1}^{T-1} \gamma(j,t)\Big(x(t) - m_x(j,t)\Big)x(t-k) = 0, \tag{3}$$

$$\sum_{t}^{T-1} \gamma(j,t)\Big(x(t) - m_x(j,t)\Big) = 0, \tag{4}$$

Here, $\gamma(j,t)$ is the posterior probability of the states, and

$$m_x(j,t) = \sum_{k=1}^{p} g(k)x(t-k) + m_r(j), \tag{5}$$

where $m_r(j)$ is the mean of state j.

The state posterior $\gamma(j,t)$ is estimated by a forward-backward induction, based on an initial estimate of the AR coefficients. The LS estimates of the AR coefficients are used as the initialization. With the voiced speech signal as an example, the voiced speech can be modeled as a noisy impulse train filtered by a vocal tract filter, and a two-state HMM is sufficient for representing the impulse train: a state with a mean equal to the magnitude of the impulses, and a state with a zero mean. For a frame that does not contain the onset of the impulse train, there must be ripples, or ringing, at the beginning of the frame, which is originated from an impulse in the previous frame. If the ringing is large enough, it will be erroneously interpreted by the algorithm as having a non-zero-mean state at the beginning of the frame although the true state is a zero-mean state. The wrong decision on the state certainly has a negative impact on the subsequent estimation of parameters. To illustrate the problem, in Fig. 1, we plot the log-spectral distance (LSD) between an estimated spectrum and the true spectrum for frames of signal beginning at different time instants. The signal is a synthetic speech signal, generated by filtering a noisy impulse train with a 10th order AR filter (the first 200 samples of the signal and its excitation are shown in Fig. 2). The first impulse, i.e. the onset, is located at the 50th sample. A hundred frames with length of 320

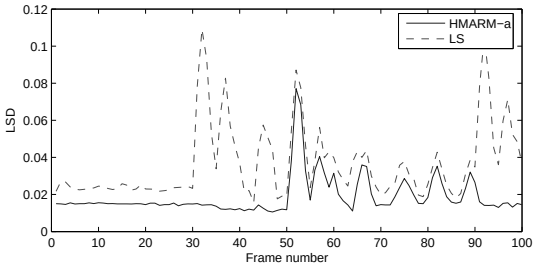


Figure 1: The log-spectral distances between the true AR spectrum and the estimates.

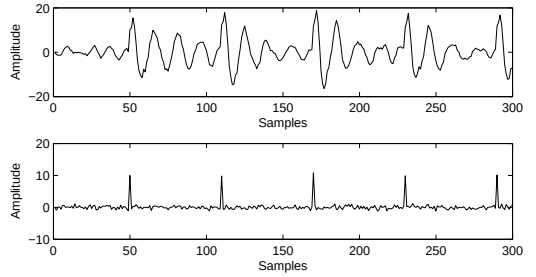


Figure 2: The synthetic signal waveform (upper panel) and its excitation (lower panel).

samples are taken from the signal by shifting the frame one sample each time. The figure shows that for the first 50 frames, i.e. all the frames that contain the onset, the spectral distortions of the HMARM spectra are low and constant. In the rest 50 frames, where the onset impulse is absent, the distortion is generally much higher. Also shown in Fig. 1 are the distortion curve for the LS spectral estimates. These curves show that the problem with the LS method is of another kind, which was pointed out in [1].

The results of the HMARM shown in Fig. 1 are without any preprocessing. To avoid the problem, in [1] and [2], a preprocessor detects the position of the first impulse of the excitation in the current frame, and sets all samples before this position to zero, such that large ripples trailing from the previous frame are removed. The problem with this solution is that removing samples reduces data efficiency of the algorithm. The reliability of the impulse detector is also a concern. Another solution is to calculate the ripples from the previous frame, using the estimated AR filter and the impulses of the previous frame, and subtract it from the current frame. This solution also reduces data efficiency, since a certain part of the signal energy is discarded, which could have been used by the estimator. Furthermore, the ringing will be subtracted using an inaccurate estimate of the AR coefficients. Moreover, these solutions add extra complexity to the algorithm.

The solution we propose in this paper is based on the observations that the HMARM has a built in linear predictor, i.e. (5), and that an AR(p) process is a Markovian process with vector states of p-dimension. So, instead of calculate the long trailing ripples from the previous frame using estimated parameters and subtract it from the following frames, it is better to initialize the predictor of the current frame with the p samples in the end of the previous frame, which gives the state estimator all the information about the past. therefore the causality problem is avoided.

To implement this solution, we only have to change the way the data matrix and the p covariance vectors are populated. They are used in the matrix form of the predictor (5) and the equations system (**??**) in the following forms:

$$\begin{bmatrix} x_0 & x_{-1} & x_{-2} & \cdots & x_{-p+1} \\ x_1 & x_0 & x_{-1} & \cdots & x_{-p+2} \\ x_2 & x_1 & x_0 & \cdots & x_{-p+3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{T-1} & x_{T-2} & x_{T-3} & \cdots & x_{T-p} \end{bmatrix}, \tag{6}$$

where T is the frame length, and

$$\begin{bmatrix} x_1 x_{1-k}, & x_2 x_{2-k}, & \cdots, & x_T x_{T-k} \end{bmatrix}^t, k = 1, \cdots, p. \tag{7}$$

In the frame based implementation the samples with negative indices are of value zero. To provide the estimator a correct starting state, the samples in the previous frame must be put into the appropriate positions of the matrices. In the case that the previous frame is missing, the first p rows of the matrices in (6) and (7) must be removed, so that there is no un-populated elements (the zeros) in the matrices. This is formally similar to the covariance method of the LS analysis of AR models [3]. Therefore, we term it the covariance method HMARM, and the original implementation the autocorrelation method HMARM. The LSD of the two implementations are plotted in Fig. 3 for comparison. It is clear from this figure that the covariance method HMARM maintains its good performance for all frames. Notice that for frames that contain the onset impulse, the performance of the covariance method HMARM is similar to the autocorrelation method HMARM. This is in contrast to the LS, whose covariance method always outperforms its autocorrelation method, given that the signal length is small.

3 HMARM for spectral analysis

Now, we discuss some properties of the HMARM that can be beneficial in the AR spectral analysis. The HMARM hereafter refers to the covariance method implementation.

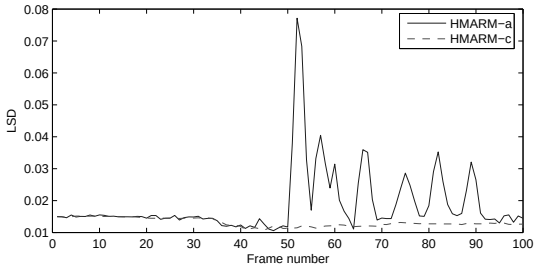


Figure 3: The log-spectral distances between the true AR spectrum and the estimates. HMARM-a: the autocorrelation method of the HMARM; HMARM-c: the covariance method of the HMARM.

3.1 Window design and covariance methods

As shown in [1] and [2], the HMARM estimate of the AR spectrum has significantly lower bias and variance than the LPC analysis, which is an autocorrelation LS method. The variance studied therein is the shift variance, where the set of realizations of an AR process is generated by shifting a time window many times with one sample as the shift step length. Other known methods for reducing the shift variance of the LS analysis are the window design and the covariance method LS. In [1], it has been shown that applying a Hamming window reduces the shift variance of the LPC analysis, but the reduced variance is still significantly larger than that of the HMARM. Besides, any window other than the rectangular window has the side effect of reduced spectral resolutions. Here, we discuss the covariance method LS analysis, and compare the three methods under a more general variance analysis.

The covariance method LS reduces the shift variance by avoiding the boundary effect. This is done by feeding a number of samples preceding the current frame to the data matrix. In this way, the covariance matrix of the signal becomes non-Toeplitz, and thus the assumption of the signal being stationary is avoided, whereas it is still based on the assumption that the excitation is white stationary Gaussian. Therefore, for the signals of interest in this work, the large variance caused by the mismatch between the assumption and the signal is still there. To reveal a more general statistics than only the shift variance, we let the sliding window shift so many times that the beginning frames and the ending frames contain completely different samples. In this way, it is possible to show a variance consisting of both the shift variance and the variance due to different realizations. We investigate the statistical properties of the three estimators, with a synthetic speech signal and a bipolar signal received through an AR channel. The synthetic speech signal is the one used in the previous example (Fig. 2), and the received bipolar signal is generated by filtering a random [-1,1] sequence with an AR filter. They are the two typical non-Gaussian AR signals with different characteristics: the excitation of the speech signal is spectrally colored due to the periodic impulses, and has a Gaussian

	Speech		Bipolar	
	bias	variance	bias	variance
HMARM-c	0.0861	27.68	8.8×10^{-15}	4.7×10^{-24}
LS-c	0.1524	169.39	0.1595	190.41
LS-a-w	0.1276	185.90	0.1862	560.95
LS-a	0.1879	179.22	0.3100	160.46

Table 1: Comparison of biases and variances. HMARM-c: the covariance method HMARM, LS-c: the covariance method LS, LS-a-w: the autocorrelation method LS with Hamming window, LS-a: the autocorrelation method LS.

component due to the noise; while the transmitted bipolar sequence is spectrally white, and very non-Gaussian since there is no Gaussian noise in it. Tab. 1 shows the biases and variances of the three methods. The statistics are obtained from estimating 600 frames of an AR process, and the frames are obtained by moving a 320-sample window 600 times by one sample each time.

The results show that: 1) the HMARM has a far smaller variance than the autocorrelation method LS, especially for the signal that has no Gaussian componets, and 2) generally, the Hamming windowing and the covariance method do not reduce the variance of an LS AR analysis.

3.2 Avoiding spectral sampling effect

Having a more sophisticated model for the excitation makes the estimation accuracy of the HMARM superior to the traditional Gaussian AR model when applied to spectral analysis of certain non-Gaussian signals. This is because the excitation to an AR filter is often not spectrally white and/or non-Gaussian. With the HMARM, correlation in the excitation can be separated from that caused by the AR filter. Thus the estimates of the AR spectral envelop are not affected by the excitation. An example of related problems for the Gaussian AR model is the spectral sampling effect due to the impulse train structure in voiced speech.

A voiced speech signal is commonly modeled by AR filtering of an impulse train. The impulse train has a comb-shape spectrum. Although the LPC analysis is intended for estimating the spectral envelop of the signal, which models the vocal tract resonance property, the comb-shape excitation spectrum has a spectral sampling effect on the estimated spectral envelop. This causes the following problems. Firstly, when a formant peak happens to locate at one of the harmonic frequencies of the impulse train, the estimated spectral envelop will have an abnormally sharp peak. This is a well known problem for the LPC analysis in speech coding, especially for high pitch speech [4] [5]. Secondly, in the case that the formant peaks do not locate at a harmonic frequency, the peaks of the estimated spectral envelope tend to drift to the neighboring harmonic frequencies. This effect is undesired in applications such as speech synthesis

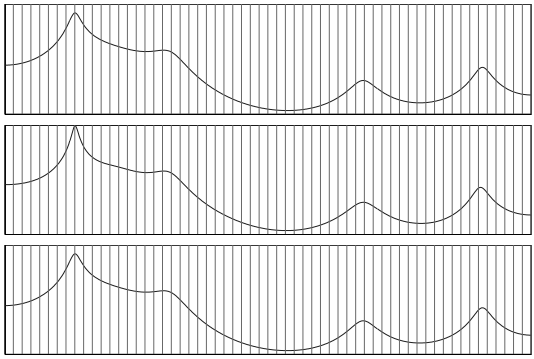


Figure 4: The AR spectra estimated by the HMARM (upper) and the LPC (middle), and the true spectrum (lower). The vertical bars show the harmonic frequencies. The pitch frequency is 133Hz.

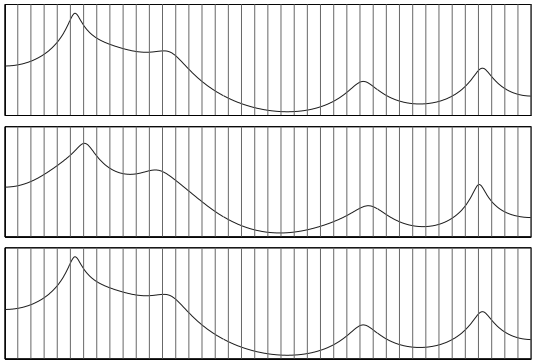


Figure 5: The AR spectra estimated by the HMARM (upper) and the LPC (middle), and the true spectrum (lower). The vertical bars show the harmonic frequencies. The pitch frequency is 200Hz.

and prosody manipulation. We compare the spectral envelopes estimated by the LPC and the HMARM, using two synthetic speech signals with pitch frequencies of 133Hz and 200Hz. Fig. 4 shows that the LPC spectral envelope has an abnormally sharp peak, while the HMARM estimate does not have the problem. Fig. 5 shows that the spectral peaks of the LPC estimate drift towards the harmonic frequencies, while the HMARM estimate has the peaks in correct positions.

3.3 Avoiding over training

Another problem associated with parametric modeling is known as over training, or over fitting. In the specific case of AR spectral analysis, over training is referred to the phenomena that when modeling the signal with a model order larger than the true order, the AR spectrum tends to fit to the FFT spectrum instead of the spectral envelope. Here

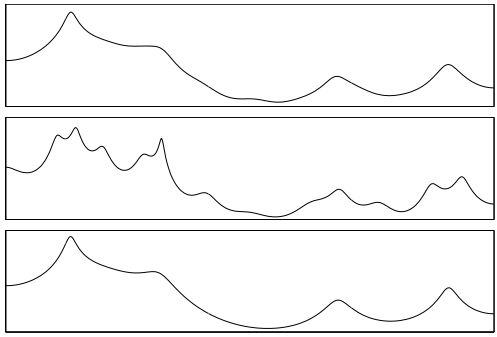


Figure 6: The AR spectra estimated by the HMARM (upper) and the LPC (middle) with order 40, and the true spectrum of order 10 (lower).

we take the bipolar signal as an example. The transmitted signal is a randomly generated bipolar signal with a white spectrum. The signal is convolved by an AR channel before it is received. The receiver tries to de-convolve the channel distortion by first estimating the channel. In general, the model order is unknown, and using a larger model order could risk over training. In Fig. 6 we show that the HMARM largely avoids the effect of over training, while the LPC spectral envelope starts representing the random peaks due to the spectrum of the transmitted signal.

4 Conclusion

In this paper, we propose a covariance method type implementation of the HMARM system identification algorithm. The method solves the causality problem that can cause the state estimator to fail in a frame based HMARM analysis. The proposed method costs no additional complexity to the system, and is proven by extensive experiments to be highly reliable. Based on the results of the covariance implementation, a few interesting issues concerning the AR spectral analysis are addressed. Examples are given for speech and digitally modulated signals with promising results.

References

[1] C. Li and S. V. Andersen, "Blind identification of non-Gaussian Autoregressive models for efficient analysis of speech signals," *Proceedings of ICASSP*, 2006.

[2] ——, "Efficient blind identification of non-Gaussian Autoregressive models with HMM modeling of the excitation," *IEEE Trans. on Signal Processing*, 2006, accepted for publication.

[3] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Prentice Hall, 2005.

[4] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, “Spectral envelope estimation and regularization,” *Proc. ICASSP*, pp. I245–I248, 2006.

[5] M. N. Murthi, “Regularized linear prediction all-pole models,” *Proc. IEEE Workshop on Speech Coding*, pp. 96–98, 2000.